data_cleaning

July 29, 2025

1 Introduction

2 Importing required libraries

[1]:
```
import numpy as np
import pandas as pd
import re
import matplotlib.pyplot as plt
import seaborn as sns
import dask.dataframe as dd
import glob
```

3 Check and review each csv

[2]:
```
def eval_df(dataframe):
    # Display basic info
    print("\n DATA TYPES & MISSING VALUES")
    print(dataframe.info())

    # Check for missing values
    missing_values = dataframe.isnull().sum()
    print("\n MISSING VALUES PER COLUMN")
    print(missing_values[missing_values>0])

    # Check for duplicate rows
    duplicates = dataframe.duplicated().sum()
    print(f"\n DUPLICATE ROWS FOUND: {duplicates}")

    # Display basic statistics
    print("\n SUMMARY STATISTICS")
    print(dataframe.describe(include="all"))
```

3.1 Country stat

[3]:
```
# Start the dataframe with the a list of country and country code
country_stat = pd.read_csv("data/country_codes.csv")
country_stat.head()
```

[3]:
```
           Country Code
0      Afghanistan  AFG
1    Aland Islands  ALA
2          Albania  ALB
3          Algeria  DZA
4   American Samoa  ASM
```

[4]:
```
folder_path = 'data/country_stat/'
csv_files = glob.glob(folder_path + "*.csv")

for file in csv_files:
    df = pd.read_csv(file)
    df.drop(columns=["Country"], inplace=True)
    country_stat = pd.merge(country_stat, df, on = "Code", how = "outer")

eval_df(country_stat)
```

```
 DATA TYPES & MISSING VALUES
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 257 entries, 0 to 256
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Country                      251 non-null    object
 1   Code                         252 non-null    object
 2   median_age                   201 non-null    float64
 3   gdp_per_capita               164 non-null    float64
 4   land_boundaries              246 non-null    float64
 5   coastline                    244 non-null    float64
 6   num_border_countries         244 non-null    float64
 7   border_countries             162 non-null    object
 8   political_regime             174 non-null    object
 9   hospital_beds_per_1000       160 non-null    float64
 10  corruption_perception_index  180 non-null    float64
 11  unemployment                 187 non-null    float64
 12  land_area_sqkm               195 non-null    float64
 13  population_density           200 non-null    float64
 14  urban_population             197 non-null    float64
 15  poverty                      147 non-null    float64
 16  gini_index                   147 non-null    float64
dtypes: float64(13), object(4)
```

```
memory usage: 34.3+ KB
None

 MISSING VALUES PER COLUMN
Country                          6
Code                             5
median_age                      56
gdp_per_capita                  93
land_boundaries                 11
coastline                       13
num_border_countries            13
border_countries                95
political_regime                83
hospital_beds_per_1000          97
corruption_perception_index     77
unemployment                    70
land_area_sqkm                  62
population_density              57
urban_population                60
poverty                        110
gini_index                     110
dtype: int64

 DUPLICATE ROWS FOUND: 0

 SUMMARY STATISTICS
       Country Code  median_age  gdp_per_capita  land_boundaries  \
count      251  252  201.000000      164.000000       246.000000
unique     251  252         NaN             NaN              NaN
top      Aruba  ABW         NaN             NaN              NaN
freq         1    1         NaN             NaN              NaN
mean       NaN  NaN   29.041841    19262.506680      2199.288618
std        NaN  NaN    9.253179    20497.291382      3334.156438
min        NaN  NaN   14.368000      623.559800         0.000000
25%        NaN  NaN   20.903000     4498.190625         0.000000
50%        NaN  NaN   28.361000    12515.447500       884.000000
75%        NaN  NaN   36.543000    27700.940250      3211.500000
max        NaN  NaN   54.642000   150732.220000     22457.000000

            coastline  num_border_countries  \
count      244.000000            244.000000
unique            NaN                   NaN
top               NaN                   NaN
freq              NaN                   NaN
mean      3262.705738              2.590164
std      14417.466167              2.620759
min          0.000000              0.000000
25%         54.750000              0.000000
```

50%	387.500000	2.000000
75%	1821.250000	4.000000
max	202080.000000	14.000000

	border_countries
count	162
unique	162
top	China 91 km; Iran 921 km; Pakistan 2,670 km; T…
freq	1
mean	NaN
std	NaN
min	NaN
25%	NaN
50%	NaN
75%	NaN
max	NaN

	political_regime	hospital_beds_per_1000
count	174	160.000000
unique	4	NaN
top	electoral_autocracies	NaN
freq	61	NaN
mean	NaN	3.016125
std	NaN	2.713317
min	NaN	0.170000
25%	NaN	1.165000
50%	NaN	2.395000
75%	NaN	4.140000
max	NaN	22.020000

	corruption_perception_index	unemployment	land_area_sqkm
count	180.000000	187.000000	1.950000e+02
unique	NaN	NaN	NaN
top	NaN	NaN	NaN
freq	NaN	NaN	NaN
mean	43.166667	7.293941	6.651783e+05
std	18.960264	5.673796	1.827517e+06
min	9.000000	0.100000	2.084000e+00
25%	29.000000	3.448000	2.392500e+04
50%	39.500000	5.206000	1.204100e+05
75%	56.000000	10.334500	5.194300e+05
max	87.000000	28.468000	1.637687e+07

	population_density	urban_population	poverty	gini_index
count	200.000000	197.000000	147.000000	147.000000
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN

```
mean            294.578268         59.427223   10.659874    0.367931
std            1414.506600         23.161899   17.662905    0.075427
min               0.136699         13.250000    0.000000    0.232323
25%              31.278362         41.612000    0.255667    0.311884
50%              84.848432         60.308000    1.341573    0.353599
75%             209.366408         78.099000   15.344651    0.415596
max           18297.025000        100.000000   78.942020    0.630258
```

[5]:
```
country_stat.query("Country.isna() or Country == ''")
```

[5]:
```
    Country Code  median_age  gdp_per_capita  land_boundaries  coastline  \
42      NaN  CHI         NaN             NaN              NaN        NaN
252     NaN  NaN         NaN             NaN            156.0       83.8
253     NaN  NaN         NaN             NaN              0.0       74.1
254     NaN  NaN         NaN             NaN              0.0       11.1
255     NaN  NaN         NaN             NaN              0.0      124.1
256     NaN  NaN         NaN             NaN              0.0      518.0

    num_border_countries border_countries political_regime  \
42                   NaN              NaN              NaN
252                  1.0    Cyprus 156 km              NaN
253                  0.0              NaN              NaN
254                  0.0              NaN              NaN
255                  0.0              NaN              NaN
256                  0.0              NaN              NaN

    hospital_beds_per_1000  corruption_perception_index  unemployment  \
42                     NaN                          NaN         6.408
252                    NaN                          NaN           NaN
253                    NaN                          NaN           NaN
254                    NaN                          NaN           NaN
255                    NaN                          NaN           NaN
256                    NaN                          NaN           NaN

    land_area_sqkm  population_density  urban_population  poverty  gini_index
42             NaN                 NaN               NaN      NaN         NaN
252            NaN                 NaN               NaN      NaN         NaN
253            NaN                 NaN               NaN      NaN         NaN
254            NaN                 NaN               NaN      NaN         NaN
255            NaN                 NaN               NaN      NaN         NaN
256            NaN                 NaN               NaN      NaN         NaN
```

[6]:
```
country_stat = country_stat[~(country_stat['Country'].isna() |␣
 ↪(country_stat['Country']==''))]
eval_df(country_stat)
```

```
DATA TYPES & MISSING VALUES
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 251 entries, 0 to 251
Data columns (total 17 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Country                       251 non-null    object
 1   Code                          251 non-null    object
 2   median_age                    201 non-null    float64
 3   gdp_per_capita                164 non-null    float64
 4   land_boundaries               241 non-null    float64
 5   coastline                     239 non-null    float64
 6   num_border_countries          239 non-null    float64
 7   border_countries              161 non-null    object
 8   political_regime              174 non-null    object
 9   hospital_beds_per_1000        160 non-null    float64
 10  corruption_perception_index   180 non-null    float64
 11  unemployment                  186 non-null    float64
 12  land_area_sqkm                195 non-null    float64
 13  population_density            200 non-null    float64
 14  urban_population              197 non-null    float64
 15  poverty                       147 non-null    float64
 16  gini_index                    147 non-null    float64
dtypes: float64(13), object(4)
memory usage: 35.3+ KB
None

 MISSING VALUES PER COLUMN
median_age                     50
gdp_per_capita                 87
land_boundaries                10
coastline                      12
num_border_countries           12
border_countries               90
political_regime               77
hospital_beds_per_1000         91
corruption_perception_index    71
unemployment                   65
land_area_sqkm                 56
population_density             51
urban_population               54
poverty                       104
gini_index                    104
dtype: int64

 DUPLICATE ROWS FOUND: 0

 SUMMARY STATISTICS
      Country Code  median_age  gdp_per_capita  land_boundaries  \
```

count	251	251	201.000000	164.000000	241.000000
unique	251	251	NaN	NaN	NaN
top	Aruba	ABW	NaN	NaN	NaN
freq	1	1	NaN	NaN	NaN
mean	NaN	NaN	29.041841	19262.506680	2244.269710
std	NaN	NaN	9.253179	20497.291382	3353.826588
min	NaN	NaN	14.368000	623.559800	0.000000
25%	NaN	NaN	20.903000	4498.190625	0.000000
50%	NaN	NaN	28.361000	12515.447500	1066.000000
75%	NaN	NaN	36.543000	27700.940250	3261.000000
max	NaN	NaN	54.642000	150732.220000	22457.000000

	coastline	num_border_countries	\
count	239.000000	239.000000	
unique	NaN	NaN	
top	NaN	NaN	
freq	NaN	NaN	
mean	3327.569456	2.640167	
std	14561.021117	2.624268	
min	0.000000	0.000000	
25%	53.500000	0.000000	
50%	400.000000	2.000000	
75%	1918.000000	4.000000	
max	202080.000000	14.000000	

	border_countries	\
count	161	
unique	161	
top	China 91 km; Iran 921 km; Pakistan 2,670 km; T…	
freq	1	
mean	NaN	
std	NaN	
min	NaN	
25%	NaN	
50%	NaN	
75%	NaN	
max	NaN	

	political_regime	hospital_beds_per_1000	\
count	174	160.000000	
unique	4	NaN	
top	electoral_autocracies	NaN	
freq	61	NaN	
mean	NaN	3.016125	
std	NaN	2.713317	
min	NaN	0.170000	
25%	NaN	1.165000	
50%	NaN	2.395000	

```
75%                          NaN                4.140000
max                          NaN               22.020000

        corruption_perception_index  unemployment  land_area_sqkm  \
count                    180.000000    186.000000    1.950000e+02
unique                          NaN           NaN             NaN
top                             NaN           NaN             NaN
freq                            NaN           NaN             NaN
mean                      43.166667      7.298704    6.651783e+05
std                       18.960264      5.688735    1.827517e+06
min                        9.000000      0.100000    2.084000e+00
25%                       29.000000      3.433500    2.392500e+04
50%                       39.500000      5.182500    1.204100e+05
75%                       56.000000     10.361750    5.194300e+05
max                       87.000000     28.468000    1.637687e+07

        population_density  urban_population     poverty  gini_index
count           200.000000        197.000000  147.000000  147.000000
unique                 NaN               NaN         NaN         NaN
top                    NaN               NaN         NaN         NaN
freq                   NaN               NaN         NaN         NaN
mean            294.578268         59.427223   10.659874    0.367931
std            1414.506600         23.161899   17.662905    0.075427
min               0.136699         13.250000    0.000000    0.232323
25%              31.278362         41.612000    0.255667    0.311884
50%              84.848432         60.308000    1.341573    0.353599
75%             209.366408         78.099000   15.344651    0.415596
max           18297.025000        100.000000   78.942020    0.630258
```

[7]:
```
missing_df = country_stat.copy()
# since there are island countries that are islands and has no border countries␣
 ↪I dropped the border countries for now, to narrow down the geographic data I␣
 ↪would just use land_boundaries
missing_df.drop(columns=['border_countries','coastline',␣
 ↪'num_border_countries'], inplace=True)
missing_df['missing_count'] = missing_df.isna().sum(axis=1)
missing_df = missing_df[missing_df['missing_count']>0].
 ↪sort_values(by='missing_count', ascending=False)
missing_df.style.set_table_attributes('style="height:300px; overflow-y:scroll;␣
 ↪display:block;"')
```

[7]:
```
<pandas.io.formats.style.Styler at 0x78b9e4380ad0>
```

[8]:
```
eval_df(missing_df)
```

```
 DATA TYPES & MISSING VALUES
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 141 entries, 4 to 250
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Country                       141 non-null    object
 1   Code                          141 non-null    object
 2   median_age                    91 non-null     float64
 3   gdp_per_capita                54 non-null     float64
 4   land_boundaries               131 non-null    float64
 5   political_regime              64 non-null     object
 6   hospital_beds_per_1000        50 non-null     float64
 7   corruption_perception_index   70 non-null     float64
 8   unemployment                  76 non-null     float64
 9   land_area_sqkm                85 non-null     float64
 10  population_density            90 non-null     float64
 11  urban_population              87 non-null     float64
 12  poverty                       37 non-null     float64
 13  gini_index                    37 non-null     float64
 14  missing_count                 141 non-null    int64
dtypes: float64(11), int64(1), object(3)
memory usage: 17.6+ KB
None

 MISSING VALUES PER COLUMN
median_age                     50
gdp_per_capita                 87
land_boundaries                10
political_regime               77
hospital_beds_per_1000         91
corruption_perception_index    71
unemployment                   65
land_area_sqkm                 56
population_density             51
urban_population               54
poverty                       104
gini_index                    104
dtype: int64

 DUPLICATE ROWS FOUND: 0

 SUMMARY STATISTICS
             Country Code  median_age  gdp_per_capita  land_boundaries  \
count            141  141   91.000000       54.000000       131.000000
unique           141  141         NaN             NaN              NaN
top     Aland Islands  ALA         NaN             NaN              NaN
freq               1    1         NaN             NaN              NaN
mean             NaN  NaN   27.212308    16204.828681      1214.875954
std              NaN  NaN    8.816120    17814.175606      2227.925018
```

min	NaN	NaN	14.368000	623.559800	0.000000
25%	NaN	NaN	20.000500	3514.452975	0.000000
50%	NaN	NaN	26.776000	10276.456000	0.000000
75%	NaN	NaN	32.991000	17750.520750	1552.000000
max	NaN	NaN	54.642000	74649.734000	11968.000000

	political_regime	hospital_beds_per_1000	\
count	64	50.000000	
unique	4	NaN	
top	electoral_autocracies	NaN	
freq	27	NaN	
mean	NaN	3.140200	
std	NaN	3.417469	
min	NaN	0.170000	
25%	NaN	1.427500	
50%	NaN	2.575000	
75%	NaN	3.707500	
max	NaN	22.020000	

	corruption_perception_index	unemployment	land_area_sqkm	\
count	70.000000	76.000000	8.500000e+01	
unique	NaN	NaN	NaN	
top	NaN	NaN	NaN	
freq	NaN	NaN	NaN	
mean	38.528571	8.681934	3.309309e+05	
std	18.169668	6.708543	5.912430e+05	
min	9.000000	0.119000	2.084000e+00	
25%	25.000000	3.177250	8.100000e+02	
50%	35.500000	5.756000	3.814000e+04	
75%	51.000000	12.332000	4.104500e+05	
max	87.000000	28.468000	2.736690e+06	

	population_density	urban_population	poverty	gini_index	\
count	90.000000	87.000000	37.000000	37.000000	
unique	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	
mean	470.807303	56.655149	18.695479	0.384675	
std	2087.756420	25.583067	23.433066	0.088113	
min	0.136699	13.250000	0.000000	0.270958	
25%	24.061025	35.666000	0.945762	0.321253	
50%	81.988720	54.835000	9.963333	0.356902	
75%	243.551045	79.167500	26.062035	0.438164	
max	18297.025000	100.000000	78.942020	0.630258	

	missing_count
count	141.000000
unique	NaN

```
top              NaN
freq             NaN
mean        5.815603
std         4.240288
min         1.000000
25%         2.000000
50%         4.000000
75%        11.000000
max        12.000000
```

I would be conservative and would just only remove all countries that only has one variable

[9]:
```
missing_df = missing_df[missing_df['missing_count']>=11].
 ↪sort_values(by='missing_count', ascending=False)
missing_df
```

[9]:

	Country	Code	median_age	\
4	Aland Islands	ALA	NaN	
20	Bonaire, Sint Eustatius and Saba	BES	NaN	
159	Mayotte	MYT	NaN	
188	Reunion	REU	NaN	
86	Guadeloupe	GLP	NaN	
155	Martinique	MTQ	NaN	
247	Turkish Rep N Cyprus (temporary code)	XTX	NaN	
27	Saint Barthélemy	BLM	NaN	
11	Antarctica	ATA	NaN	
12	French Southern Territories	ATF	NaN	
0	Aruba	ABW	NaN	
10	American Samoa	ASM	NaN	
3	Anguilla	AIA	NaN	
154	Montserrat	MSR	NaN	
244	Wallis and Futuna	WLF	NaN	
233	United States Minor Outlying Islands	UMI	NaN	
213	Sint Maarten	SXM	NaN	
237	Holy See (Vatican City State)	VAT	NaN	
240	Virgin Islands, British	VGB	NaN	
40	Cocos (Keeling) Islands	CCK	NaN	
30	Bermuda	BMU	NaN	
75	Falkland Islands (Malvinas)	FLK	NaN	
77	Faroe Islands	FRO	NaN	
84	Gibraltar	GIB	NaN	
82	Guernsey	GGY	NaN	
49	Cook Islands	COK	NaN	
57	Cayman Islands	CYM	NaN	
56	Christmas Island	CXR	NaN	
36	Bouvet Island	BVT	NaN	
106	British Indian Ocean Territory	IOT	NaN	
204	Saint Pierre and Miquelon	SPM	NaN	

```
216                   Turks and Caicos Islands  TCA         NaN
221                                    Tokelau  TKL         NaN
98                 Heard and Mc Donald Islands  HMD         NaN
104                                Isle of Man  IMN         NaN
196  South Georgia & The South Sandwich Islands  SGS         NaN
197                                Saint Helena  SHN         NaN
163                              Norfolk Island  NFK         NaN
175                                   Pitcairn  PCN         NaN
198                      Svalbard and Jan Mayen  SJM         NaN
166                                        Niue  NIU         NaN
114                                      Jersey  JEY         NaN
55                                     Curaçao  CUW         NaN

     gdp_per_capita  land_boundaries political_regime  hospital_beds_per_1000  \
4               NaN              NaN              NaN                     NaN
20              NaN              NaN              NaN                     NaN
159             NaN              NaN              NaN                     NaN
188             NaN              NaN              NaN                     NaN
86              NaN              NaN              NaN                     NaN
155             NaN              NaN              NaN                     NaN
247             NaN              NaN              NaN                     NaN
27              NaN              0.0              NaN                     NaN
11              NaN              0.0              NaN                     NaN
12              NaN              0.0              NaN                     NaN
0               NaN              0.0              NaN                     NaN
10              NaN              0.0              NaN                     NaN
3               NaN              0.0              NaN                     NaN
154             NaN              0.0              NaN                     NaN
244             NaN              0.0              NaN                     NaN
233             NaN              0.0              NaN                     NaN
213             NaN             16.0              NaN                     NaN
237             NaN              3.4              NaN                     NaN
240             NaN              0.0              NaN                     NaN
40              NaN              0.0              NaN                     NaN
30              NaN              0.0              NaN                     NaN
75              NaN              0.0              NaN                     NaN
77              NaN              0.0              NaN                     NaN
84              NaN              1.2              NaN                     NaN
82              NaN              0.0              NaN                     NaN
49              NaN              0.0              NaN                     NaN
57              NaN              0.0              NaN                     NaN
56              NaN              0.0              NaN                     NaN
36              NaN              0.0              NaN                     NaN
106             NaN              0.0              NaN                     NaN
204             NaN              0.0              NaN                     NaN
216             NaN              0.0              NaN                     NaN
221             NaN              0.0              NaN                     NaN
```

```
98              NaN              0.0              NaN                      NaN
104             NaN              0.0              NaN                      NaN
196             NaN              0.0              NaN                      NaN
197             NaN              0.0              NaN                      NaN
163             NaN              0.0              NaN                      NaN
175             NaN              0.0              NaN                      NaN
198             NaN              0.0              NaN                      NaN
166             NaN              0.0              NaN                      NaN
114             NaN              0.0              NaN                      NaN
55              NaN              0.0              NaN                      NaN

     corruption_perception_index  unemployment  land_area_sqkm  \
4                            NaN           NaN             NaN
20                           NaN           NaN             NaN
159                          NaN           NaN             NaN
188                          NaN           NaN             NaN
86                           NaN           NaN             NaN
155                          NaN           NaN             NaN
247                          NaN           NaN             NaN
27                           NaN           NaN             NaN
11                           NaN           NaN             NaN
12                           NaN           NaN             NaN
0                            NaN           NaN             NaN
10                           NaN           NaN             NaN
3                            NaN           NaN             NaN
154                          NaN           NaN             NaN
244                          NaN           NaN             NaN
233                          NaN           NaN             NaN
213                          NaN           NaN             NaN
237                          NaN           NaN             NaN
240                          NaN           NaN             NaN
40                           NaN           NaN             NaN
30                           NaN           NaN             NaN
75                           NaN           NaN             NaN
77                           NaN           NaN             NaN
84                           NaN           NaN             NaN
82                           NaN           NaN             NaN
49                           NaN           NaN             NaN
57                           NaN           NaN             NaN
56                           NaN           NaN             NaN
36                           NaN           NaN             NaN
106                          NaN           NaN             NaN
204                          NaN           NaN             NaN
216                          NaN           NaN             NaN
221                          NaN           NaN             NaN
98                           NaN           NaN             NaN
104                          NaN           NaN             NaN
```

196	NaN	NaN	NaN	
197	NaN	NaN	NaN	
163	NaN	NaN	NaN	
175	NaN	NaN	NaN	
198	NaN	NaN	NaN	
166	NaN	NaN	NaN	
114	NaN	NaN	NaN	
55	NaN	NaN	NaN	

	population_density	urban_population	poverty	gini_index	missing_count
4	NaN	NaN	NaN	NaN	12
20	NaN	NaN	NaN	NaN	12
159	NaN	NaN	NaN	NaN	12
188	NaN	NaN	NaN	NaN	12
86	NaN	NaN	NaN	NaN	12
155	NaN	NaN	NaN	NaN	12
247	NaN	NaN	NaN	NaN	12
27	NaN	NaN	NaN	NaN	11
11	NaN	NaN	NaN	NaN	11
12	NaN	NaN	NaN	NaN	11
0	NaN	NaN	NaN	NaN	11
10	NaN	NaN	NaN	NaN	11
3	NaN	NaN	NaN	NaN	11
154	NaN	NaN	NaN	NaN	11
244	NaN	NaN	NaN	NaN	11
233	NaN	NaN	NaN	NaN	11
213	NaN	NaN	NaN	NaN	11
237	NaN	NaN	NaN	NaN	11
240	NaN	NaN	NaN	NaN	11
40	NaN	NaN	NaN	NaN	11
30	NaN	NaN	NaN	NaN	11
75	NaN	NaN	NaN	NaN	11
77	NaN	NaN	NaN	NaN	11
84	NaN	NaN	NaN	NaN	11
82	NaN	NaN	NaN	NaN	11
49	NaN	NaN	NaN	NaN	11
57	NaN	NaN	NaN	NaN	11
56	NaN	NaN	NaN	NaN	11
36	NaN	NaN	NaN	NaN	11
106	NaN	NaN	NaN	NaN	11
204	NaN	NaN	NaN	NaN	11
216	NaN	NaN	NaN	NaN	11
221	NaN	NaN	NaN	NaN	11
98	NaN	NaN	NaN	NaN	11
104	NaN	NaN	NaN	NaN	11
196	NaN	NaN	NaN	NaN	11
197	NaN	NaN	NaN	NaN	11

```
163                 NaN               NaN       NaN          NaN              11
175                 NaN               NaN       NaN          NaN              11
198                 NaN               NaN       NaN          NaN              11
166                 NaN               NaN       NaN          NaN              11
114                 NaN               NaN       NaN          NaN              11
55                  NaN               NaN       NaN          NaN              11
```

[10]:
```
country_stat = country_stat[~country_stat['Code'].isin(missing_df['Code'])]
eval_df(country_stat)
```

```
 DATA TYPES & MISSING VALUES
<class 'pandas.core.frame.DataFrame'>
Index: 208 entries, 1 to 251
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Country                      208 non-null    object
 1   Code                         208 non-null    object
 2   median_age                   201 non-null    float64
 3   gdp_per_capita               164 non-null    float64
 4   land_boundaries              205 non-null    float64
 5   coastline                    205 non-null    float64
 6   num_border_countries         205 non-null    float64
 7   border_countries             158 non-null    object
 8   political_regime             174 non-null    object
 9   hospital_beds_per_1000       160 non-null    float64
 10  corruption_perception_index  180 non-null    float64
 11  unemployment                 186 non-null    float64
 12  land_area_sqkm               195 non-null    float64
 13  population_density           200 non-null    float64
 14  urban_population             197 non-null    float64
 15  poverty                      147 non-null    float64
 16  gini_index                   147 non-null    float64
dtypes: float64(13), object(4)
memory usage: 29.2+ KB
None

 MISSING VALUES PER COLUMN
median_age                     7
gdp_per_capita                44
land_boundaries                3
coastline                      3
num_border_countries           3
border_countries              50
political_regime              34
hospital_beds_per_1000        48
corruption_perception_index   28
```

```
unemployment                   22
land_area_sqkm                 13
population_density              8
urban_population               11
poverty                        61
gini_index                     61
dtype: int64

 DUPLICATE ROWS FOUND: 0

 SUMMARY STATISTICS
            Country Code  median_age  gdp_per_capita  land_boundaries  \
count           208  208  201.000000      164.000000       205.000000
unique          208  208         NaN             NaN              NaN
top     Afghanistan  AFG         NaN             NaN              NaN
freq              1    1         NaN             NaN              NaN
mean            NaN  NaN   29.041841    19262.506680      2638.284878
std             NaN  NaN    9.253179    20497.291382      3491.235636
min             NaN  NaN   14.368000      623.559800         0.000000
25%             NaN  NaN   20.903000     4498.190625        75.000000
50%             NaN  NaN   28.361000    12515.447500      1587.000000
75%             NaN  NaN   36.543000    27700.940250      4046.000000
max             NaN  NaN   54.642000   150732.220000     22457.000000

            coastline  num_border_countries  \
count      205.000000            205.000000
unique            NaN                   NaN
top               NaN                   NaN
freq              NaN                   NaN
mean      3731.802927              3.063415
std      15640.819111              2.599185
min          0.000000              0.000000
25%         46.600000              1.000000
50%        499.000000              3.000000
75%       2389.000000              5.000000
max     202080.000000             14.000000

                                        border_countries  \
count                                                158
unique                                               158
top     China 91 km; Iran 921 km; Pakistan 2,670 km; T…
freq                                                   1
mean                                                 NaN
std                                                  NaN
min                                                  NaN
25%                                                  NaN
50%                                                  NaN
75%                                                  NaN
```

```
max                                                    NaN

            political_regime  hospital_beds_per_1000  \
count                        174              160.000000
unique                         4                     NaN
top     electoral_autocracies                        NaN
freq                          61                     NaN
mean                         NaN                3.016125
std                          NaN                2.713317
min                          NaN                0.170000
25%                          NaN                1.165000
50%                          NaN                2.395000
75%                          NaN                4.140000
max                          NaN               22.020000

        corruption_perception_index  unemployment  land_area_sqkm  \
count                    180.000000    186.000000    1.950000e+02
unique                          NaN           NaN             NaN
top                             NaN           NaN             NaN
freq                            NaN           NaN             NaN
mean                      43.166667      7.298704    6.651783e+05
std                       18.960264      5.688735    1.827517e+06
min                        9.000000      0.100000    2.084000e+00
25%                       29.000000      3.433500    2.392500e+04
50%                       39.500000      5.182500    1.204100e+05
75%                       56.000000     10.361750    5.194300e+05
max                       87.000000     28.468000    1.637687e+07

        population_density  urban_population     poverty  gini_index
count           200.000000        197.000000  147.000000  147.000000
unique                 NaN               NaN         NaN         NaN
top                    NaN               NaN         NaN         NaN
freq                   NaN               NaN         NaN         NaN
mean            294.578268         59.427223   10.659874    0.367931
std            1414.506600         23.161899   17.662905    0.075427
min               0.136699         13.250000    0.000000    0.232323
25%              31.278362         41.612000    0.255667    0.311884
50%              84.848432         60.308000    1.341573    0.353599
75%             209.366408         78.099000   15.344651    0.415596
max           18297.025000        100.000000   78.942020    0.630258
```

3.2 National Policy

[11]:
```
national_policy = pd.read_csv('data/national_policy.csv')
national_policy.head()
```

[11]:

	Code	Country	area_effect	log_type	measure	COMMENTS	implementation_date
0	AFG	Afghanistan	national	induction	Awareness campaigns	MoPH begins announcements on their facebook to…	1/24/2020
1	AFG	Afghanistan	national	induction	Health screenings in airports and border cross…	Health teams at airports will check passengers…	1/26/2020
2	AFG	Afghanistan	national	induction	Health screenings in airports and border cross…	Health screenings of all passengers at airports.	1/27/2020
3	AFG	Afghanistan	national	induction	International flights suspension	Flights to China are suspended.	1/27/2020
4	AFG	Afghanistan	national	induction	Border checks	All China and Iran nationals	2/1/2020

[12]:
```
national_policy.describe()
```

[12]:

	Code	Country	area_effect	log_type	measure	COMMENTS	implementation_date
count	23923	23923	23923	23923	23923	23799	23636
unique	193	193	2	2	35	23281	359
top	GBR	United Kingdom	national	induction	Economic measures	APEC economies agree to keep markets open and …	3/16/2020
freq	655	655	20256	19445	2980	19	342

[13]:
```
sorted(national_policy['measure'].unique())
```

[13]:
```
['Additional health/documents requirements upon arrival',
 'Amendments to funeral and burial regulations',
 'Awareness campaigns',
 'Border checks',
 'Border closure',
 'Changes in prison-related policies',
 'Checkpoints within the country',
 'Closure of businesses and public services',
 'Complete border closure',
 'Curfews',
```

```
 'Domestic travel restrictions',
 'Economic measures',
 'Emergency administrative structures activated or established',
 'Full lockdown',
 'General recommendations',
 'Health screenings in airports and border crossings',
 'Humanitarian exemptions',
 'International flights suspension',
 'Isolation and quarantine policies',
 'Limit product imports/exports',
 'Limit public gatherings',
 'Lockdown of refugee/idp camps or other minorities',
 'Mass population testing',
 'Military deployment',
 'Obligatory medical tests not related to COVID-19',
 'Other public health measures enforced',
 'Partial lockdown',
 'Psychological assistance and medical social work',
 'Requirement to wear protective gear in public',
 'Schools closure',
 'State of emergency declared',
 'Strengthening the public health system',
 'Surveillance and monitoring',
 'Testing policy',
 'Visa restrictions']
```

There are around 36 measures that the government made. I would focus on the measures that would directly affect population mobility

[14]:
```
measures_to_focus = ['Additional health/documents requirements upon arrival',
                     'Border checks',
                     'Border closure',
                     'Checkpoints within the country',
                     'Closure of business and public services',
                     'Complete border closure',
                     'Curfews',
                     'Domestic travel restrictions',
                     'Full lockdown',
                     'Health screenings in airports and border crossings',
                     'Humanitarian exemptions',
                     'International flights suspension',
                     'Isolation and quarantine policies',
                     'Limit public gatherings',
                     'Lockdown of refugee/idp camps or other minorities',
                     'Partial lockdown',
                     'School closure']
```

[15]:
```
# select the measures to focus
national_policy = national_policy[national_policy['measure'].
 ↪isin(measures_to_focus)]
```

[16]:
```
# create a column that is a the mix between area_of effect
national_policy['measure_type'] =␣
 ↪national_policy['area_effect']+"_"+national_policy['log_type']
# drop area_effect, log_type, and comments
national_policy.drop(columns=['area_effect', 'log_type', 'COMMENTS'],␣
 ↪inplace=True)
# get unique measure per day
national_policy.drop_duplicates(inplace=True)
```

[17]:
```
# make the measure_type into a datetime
national_policy['implementation_date'] = pd.
 ↪to_datetime(national_policy['implementation_date'], errors='coerce')
# aggregate measure_type and measure by date and country
national_policy = national_policy.
 ↪groupby(['Code','Country','implementation_date']).apply(lambda g: g.
 ↪groupby('measure_type')['measure'].unique().apply(list).to_dict()).
 ↪reset_index(name='measure_dict')
national_policy
```

```
/tmp/ipykernel_15116/2220675760.py:4: DeprecationWarning: DataFrameGroupBy.apply
operated on the grouping columns. This behavior is deprecated, and in a future
version of pandas the grouping columns will be excluded from the operation.
Either pass `include_groups=False` to exclude the groupings or explicitly select
the grouping columns after groupby to silence this warning.
  national_policy =
national_policy.groupby(['Code','Country','implementation_date']).apply(lambda
g: g.groupby('measure_type')['measure'].unique().apply(list).to_dict()).reset_in
dex(name='measure_dict')
```

[17]:

	Code	Country	implementation_date	measure_dict
0	AFG	Afghanistan	2020-01-26	
1	AFG	Afghanistan	2020-01-27	
2	AFG	Afghanistan	2020-02-01	
3	AFG	Afghanistan	2020-02-02	
4	AFG	Afghanistan	2020-02-12	
…	…	…	…	
4942	ZWE	Zimbabwe	2020-05-04	
4943	ZWE	Zimbabwe	2020-05-05	
4944	ZWE	Zimbabwe	2020-05-17	
4945	ZWE	Zimbabwe	2020-05-21	
4946	ZWE	Zimbabwe	2020-10-01	

```
0     {'national_induction': ['Health screenings in …
1     {'national_induction': ['Health screenings in …
2             {'national_induction': ['Border checks']}
3     {'national_induction': ['Isolation and quarant…
4     {'local_induction': ['Isolation and quarantine…
…                                                   …
4942      {'national_induction': ['Partial lockdown']}
4943  {'national_phaseout': ['Limit public gathering…
4944  {'national_induction': ['Partial lockdown'], '…
4945  {'national_phaseout': ['Limit public gathering…
4946  {'national_induction': ['Health screenings in …

[4947 rows x 4 columns]
```

[18]:
```
national_policy['national_induction_num'] = national_policy['measure_dict'].
 ↪apply(lambda x: len(x.get('national_induction')) if x.
 ↪get('national_induction') is not None else 0)
national_policy['national_phaseout_num'] = national_policy['measure_dict'].
 ↪apply(lambda x: len(x.get('national_phaseout_num')) if x.
 ↪get('national_phaseout_num') is not None else 0)
national_policy['local_induction_num'] = national_policy['measure_dict'].
 ↪apply(lambda x: len(x.get('local_induction')) if x.get('local_induction') is␣
 ↪not None else 0)
national_policy['local_phaseout_num'] = national_policy['measure_dict'].
 ↪apply(lambda x: len(x.get('local_phaseout_num')) if x.
 ↪get('local_phaseout_num') is not None else 0)
national_policy
```

[18]:
```
     Code      Country implementation_date  \
0     AFG  Afghanistan          2020-01-26
1     AFG  Afghanistan          2020-01-27
2     AFG  Afghanistan          2020-02-01
3     AFG  Afghanistan          2020-02-02
4     AFG  Afghanistan          2020-02-12
…   …           …                  …
4942  ZWE     Zimbabwe          2020-05-04
4943  ZWE     Zimbabwe          2020-05-05
4944  ZWE     Zimbabwe          2020-05-17
4945  ZWE     Zimbabwe          2020-05-21
4946  ZWE     Zimbabwe          2020-10-01

                                          measure_dict  \
0     {'national_induction': ['Health screenings in …
1     {'national_induction': ['Health screenings in …
2             {'national_induction': ['Border checks']}
3     {'national_induction': ['Isolation and quarant…
4     {'local_induction': ['Isolation and quarantine…
```

```
…                                                     …
4942       {'national_induction': ['Partial lockdown']}
4943  {'national_phaseout': ['Limit public gathering…
4944  {'national_induction': ['Partial lockdown'], '…
4945  {'national_phaseout': ['Limit public gathering…
4946  {'national_induction': ['Health screenings in …

      national_induction_num  national_phaseout_num  local_induction_num  \
0                          1                      0                    0
1                          2                      0                    0
2                          1                      0                    0
3                          1                      0                    0
4                          1                      0                    1
…                        …                      …                    …
4942                       1                      0                    0
4943                       0                      0                    0
4944                       1                      0                    0
4945                       0                      0                    0
4946                       1                      0                    0

      local_phaseout_num
0                      0
1                      0
2                      0
3                      0
4                      0
…                    …
4942                   0
4943                   0
4944                   0
4945                   0
4946                   0

[4947 rows x 8 columns]
```

3.3 Mobility data

[19]:
```
# load csv with selected columns excluded
excluded_cols =['iso_3166_2_code', 'census_flips_code', 'place_id']
all_cols = dd.read_csv('data/global_mobility_report.csv', sample=10000).columns.
 ↪tolist()
usecols = [col for col in all_cols if col not in excluded_cols]

mobility_data = dd.read_csv('data/global_mobility_report.csv', usecols=usecols,␣
 ↪assume_missing = True)

# Filter rows that have entry for sub_region_1, sub_region_2, metro_area
```

```
# Filter date up until 2021-01-31

#mobility_data['date'] = dd.to_datetime(mobility_data['date'], errors='coerce')
#mobility_data = mobility_data[mobility_data[['sub_region_1', 'sub_region_2',␣
 ↪'metro_area']].isnull().all(axis=1) &
 #                             (mobility_data['date']<='2021-01-31')]
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
File ~/country_mobility/country-factors-mobility/venv/lib/python3.12/
 ↪site-packages/dask/backends.py:140, in CreationDispatch.register_inplace.
 ↪<locals>.decorator.<locals>.wrapper(*args, **kwargs)
    139 try:
--> 140     return func(*args, **kwargs)
    141 except Exception as e:

File ~/country_mobility/country-factors-mobility/venv/lib/python3.12/
 ↪site-packages/dask/dataframe/io/csv.py:731, in make_reader.<locals>.
 ↪read(urlpath, blocksize, lineterminator, compression, sample, sample_rows,␣
 ↪enforce, assume_missing, storage_options, include_path_column, **kwargs)
    718 def read(
    719     urlpath,
    720     blocksize="default",
   (…)    729     **kwargs,
    730 ):
--> 731     return read_pandas(
    732         reader,
    733         urlpath,
    734         blocksize=blocksize,
    735         lineterminator=lineterminator,
    736         compression=compression,
    737         sample=sample,
    738         sample_rows=sample_rows,
    739         enforce=enforce,
    740         assume_missing=assume_missing,
    741         storage_options=storage_options,
    742         include_path_column=include_path_column,
    743         **kwargs,
    744     )

File ~/country_mobility/country-factors-mobility/venv/lib/python3.12/
 ↪site-packages/dask/dataframe/io/csv.py:524, in read_pandas(reader, urlpath,␣
 ↪blocksize, lineterminator, compression, sample, sample_rows, enforce,␣
 ↪assume_missing, storage_options, include_path_column, **kwargs)
    523     sample = blocksize
--> 524 b_out = read_bytes(
    525     urlpath,
    526     delimiter=b_lineterminator,
```

```
    527     blocksize=blocksize,
    528     sample=sample,
    529     compression=compression,
    530     include_path=include_path_column,
    531     **(storage_options or {}),
    532 )
    534 if include_path_column:

File ~/country_mobility/country-factors-mobility/venv/lib/python3.12/
 ↪site-packages/dask/bytes/core.py:111, in read_bytes(urlpath, delimiter,␣
 ↪not_zero, blocksize, sample, compression, include_path, **kwargs)
    107     raise ValueError(
    108         "Cannot do chunked reads on compressed files. "
    109         "To read, set blocksize=None"
    110     )
--> 111 size = fs.info(path)["size"]
    112 if size is None:

File ~/country_mobility/country-factors-mobility/venv/lib/python3.12/
 ↪site-packages/fsspec/implementations/local.py:100, in LocalFileSystem.
 ↪info(self, path, **kwargs)
     99 path = self._strip_protocol(path)
--> 100 out = os.stat(path, follow_symlinks=False)
    101 link = stat.S_ISLNK(out.st_mode)

FileNotFoundError: [Errno 2] No such file or directory: '/home/gener/
 ↪country_mobility/country-factors-mobility/data/global_mobility_report.csv'

The above exception was the direct cause of the following exception:

FileNotFoundError                         Traceback (most recent call last)
Cell In[19], line 3
      1 # load csv with selected columns excluded
      2 excluded_cols =['iso_3166_2_code', 'census_flips_code', 'place_id']
----> 3 all_cols = dd.read_csv(                                    , sample=10000)
 ↪columns.tolist()
      4 usecols = [col for col in all_cols if col not in excluded_cols]
      6 mobility_data = dd.read_csv('data/global_mobility_report.csv',␣
 ↪usecols=usecols, assume_missing = True)

File ~/country_mobility/country-factors-mobility/venv/lib/python3.12/
 ↪site-packages/dask/backends.py:151, in CreationDispatch.register_inplace.
 ↪<locals>.decorator.<locals>.wrapper(*args, **kwargs)
    149     raise e
    150 else:
--> 151     raise exc from e
```

```
FileNotFoundError: An error occurred while calling the read_csv method␣
 ↪registered to the pandas backend.
Original Message: [Errno 2] No such file or directory: '/home/gener/
 ↪country_mobility/country-factors-mobility/data/global_mobility_report.csv'
```

[ ]:
```
# Columns to include
usecols = [
    "country_region_code", "country_region", "sub_region_1", "sub_region_2",␣
 ↪"metro_area",
    "date", "retail_and_recreation_percent_change_from_baseline",
    "grocery_and_pharmacy_percent_change_from_baseline",
    "parks_percent_change_from_baseline",
    "transit_stations_percent_change_from_baseline",
    "workplaces_percent_change_from_baseline",
    "residential_percent_change_from_baseline"
]

# Explicit dtypes for columns Dask struggles with
dtype_fix = {
    'sub_region_1': 'object',
    'sub_region_2': 'object',
    'metro_area': 'object'
}

# Load the data
df = dd.read_csv(
    'data/global_mobility_report.csv',
    usecols=usecols,
    dtype=dtype_fix,
    parse_dates=['date'],
    assume_missing=True
)

# Filter rows
missing_cols = [
    'sub_region_1',
    'sub_region_2',
    'metro_area'
]

filtered = df[
    df[missing_cols].isnull().all(axis=1) &
    (df['date'] <= '2021-01-31')
]

# Preview safely
```

```
print(filtered.head(10))
```

```
  country_region_code        country_region sub_region_1 sub_region_2  \
0                  AE  United Arab Emirates         <NA>         <NA>
1                  AE  United Arab Emirates         <NA>         <NA>
2                  AE  United Arab Emirates         <NA>         <NA>
3                  AE  United Arab Emirates         <NA>         <NA>
4                  AE  United Arab Emirates         <NA>         <NA>
5                  AE  United Arab Emirates         <NA>         <NA>
6                  AE  United Arab Emirates         <NA>         <NA>
7                  AE  United Arab Emirates         <NA>         <NA>
8                  AE  United Arab Emirates         <NA>         <NA>
9                  AE  United Arab Emirates         <NA>         <NA>

  metro_area       date  retail_and_recreation_percent_change_from_baseline  \
0       <NA> 2020-02-15                                                0.0
1       <NA> 2020-02-16                                                1.0
2       <NA> 2020-02-17                                               -1.0
3       <NA> 2020-02-18                                               -2.0
4       <NA> 2020-02-19                                               -2.0
5       <NA> 2020-02-20                                               -2.0
6       <NA> 2020-02-21                                               -3.0
7       <NA> 2020-02-22                                               -2.0
8       <NA> 2020-02-23                                               -1.0
9       <NA> 2020-02-24                                               -3.0

   grocery_and_pharmacy_percent_change_from_baseline  \
0                                                4.0
1                                                4.0
2                                                1.0
3                                                1.0
4                                                0.0
5                                                1.0
6                                                2.0
7                                                2.0
8                                                3.0
9                                                0.0

   parks_percent_change_from_baseline  \
0                                 5.0
1                                 4.0
2                                 5.0
3                                 5.0
4                                 4.0
5                                 6.0
6                                 6.0
7                                 4.0
8                                 3.0
```

```
9                                5.0

   transit_stations_percent_change_from_baseline  \
0                                            0.0
1                                            1.0
2                                            1.0
3                                            0.0
4                                           -1.0
5                                            1.0
6                                            0.0
7                                           -2.0
8                                           -1.0
9                                           -1.0

   workplaces_percent_change_from_baseline  \
0                                      2.0
1                                      2.0
2                                      2.0
3                                      2.0
4                                      2.0
5                                      1.0
6                                     -1.0
7                                      3.0
8                                      4.0
9                                      3.0

   residential_percent_change_from_baseline
0                                       1.0
1                                       1.0
2                                       1.0
3                                       1.0
4                                       1.0
5                                       1.0
6                                       1.0
7                                       1.0
8                                       1.0
9                                       1.0
```

[ ]:
```
filtered['date'].max().compute()
```

[ ]:
```
Timestamp('2021-01-31 00:00:00')
```

[ ]:
```
filtered.sample(frac=0.005, random_state=1).compute()
```

[ ]:
```
       country_region_code        country_region sub_region_1 sub_region_2  \
13905                   AO                Angola         <NA>         <NA>
7957                    AF           Afghanistan         <NA>         <NA>
333                     AE  United Arab Emirates         <NA>         <NA>
```

```
429500                  AT               Austria         <NA>         <NA>
18117                   AR             Argentina         <NA>         <NA>
…                     …                      …           …          …
493706                  YE                 Yemen         <NA>         <NA>
506376                  ZW              Zimbabwe         <NA>         <NA>
410222                  UY               Uruguay         <NA>         <NA>
431494                  VN               Vietnam         <NA>         <NA>
430406                  VE             Venezuela         <NA>         <NA>

       metro_area       date  \
13905        <NA> 2021-01-12
7957         <NA> 2020-09-19
333          <NA> 2021-01-13
429500       <NA> 2020-09-08
18117        <NA> 2020-06-24
…             …        …
493706       <NA> 2020-05-04
506376       <NA> 2020-05-12
410222       <NA> 2020-07-12
431494       <NA> 2020-08-08
430406       <NA> 2020-04-16

        retail_and_recreation_percent_change_from_baseline  \
13905                                                -6.0
7957                                                 35.0
333                                                 -18.0
429500                                                2.0
18117                                               -62.0
…                                                   …
493706                                              -14.0
506376                                              -35.0
410222                                              -37.0
431494                                              -20.0
430406                                              -59.0

        grocery_and_pharmacy_percent_change_from_baseline  \
13905                                                 0.0
7957                                                 21.0
333                                                  -1.0
429500                                               -5.0
18117                                               -21.0
…                                                   …
493706                                               -9.0
506376                                              -26.0
410222                                              -13.0
431494                                                2.0
430406                                              -34.0
```

```
        parks_percent_change_from_baseline  \
13905                                 28.0
7957                                  15.0
333                                  -31.0
429500                               112.0
18117                                -83.0
…                                    …
493706                               -17.0
506376                               -30.0
410222                               -66.0
431494                               -25.0
430406                               -50.0

        transit_stations_percent_change_from_baseline  \
13905                                            -3.0
7957                                            -11.0
333                                             -34.0
429500                                          -13.0
18117                                           -53.0
…                                               …
493706                                          -20.0
506376                                          -53.0
410222                                          -42.0
431494                                          -21.0
430406                                          -60.0

        workplaces_percent_change_from_baseline  \
13905                                     -9.0
7957                                       1.0
333                                      -20.0
429500                                   -28.0
18117                                    -27.0
…                                        …
493706                                   -14.0
506376                                   -31.0
410222                                    -8.0
431494                                    -8.0
430406                                   -53.0

        residential_percent_change_from_baseline
13905                                        4.0
7957                                         1.0
333                                          7.0
429500                                       1.0
18117                                       17.0
…                                           …
```

```
493706                                          5.0
506376                                         29.0
410222                                         11.0
431494                                         10.0
430406                                         22.0

[237 rows x 12 columns]
```

[ ]:
```
filtered.columns
```

[ ]:
```
Index(['country_region_code', 'country_region', 'sub_region_1', 'sub_region_2',
       'metro_area', 'date',
       'retail_and_recreation_percent_change_from_baseline',
       'grocery_and_pharmacy_percent_change_from_baseline',
       'parks_percent_change_from_baseline',
       'transit_stations_percent_change_from_baseline',
       'workplaces_percent_change_from_baseline',
       'residential_percent_change_from_baseline'],
      dtype='object')
```

[ ]:
```
filtered = filtered.drop(columns=['sub_region_1', 'sub_region_2', 'metro_area'])
sorted(filtered['country_region'].dropna().unique().compute())
```

[ ]:
```
['Afghanistan',
 'Angola',
 'Antigua and Barbuda',
 'Argentina',
 'Aruba',
 'Australia',
 'Austria',
 'Bahrain',
 'Bangladesh',
 'Barbados',
 'Belarus',
 'Belgium',
 'Belize',
 'Benin',
 'Bolivia',
 'Bosnia and Herzegovina',
 'Botswana',
 'Brazil',
 'Bulgaria',
 'Burkina Faso',
 'Cambodia',
 'Cameroon',
 'Canada',
 'Cape Verde',
```

```
'Chile',
'Colombia',
'Costa Rica',
'Croatia',
'Czechia',
"Côte d'Ivoire",
'Denmark',
'Dominican Republic',
'Ecuador',
'Egypt',
'El Salvador',
'Estonia',
'Fiji',
'Finland',
'France',
'Gabon',
'Georgia',
'Germany',
'Ghana',
'Greece',
'Guatemala',
'Guinea-Bissau',
'Haiti',
'Honduras',
'Hong Kong',
'Hungary',
'India',
'Indonesia',
'Iraq',
'Ireland',
'Israel',
'Italy',
'Jamaica',
'Japan',
'Jordan',
'Kazakhstan',
'Kenya',
'Kuwait',
'Kyrgyzstan',
'Laos',
'Latvia',
'Lebanon',
'Libya',
'Liechtenstein',
'Lithuania',
'Luxembourg',
'Malaysia',
```

```
'Mali',
'Malta',
'Mauritius',
'Mexico',
'Moldova',
'Mongolia',
'Morocco',
'Mozambique',
'Myanmar (Burma)',
'Namibia',
'Nepal',
'Netherlands',
'New Zealand',
'Nicaragua',
'Niger',
'Nigeria',
'North Macedonia',
'Norway',
'Oman',
'Pakistan',
'Panama',
'Papua New Guinea',
'Paraguay',
'Peru',
'Philippines',
'Poland',
'Portugal',
'Puerto Rico',
'Qatar',
'Romania',
'Russia',
'Rwanda',
'Réunion',
'Saudi Arabia',
'Senegal',
'Serbia',
'Singapore',
'Slovakia',
'Slovenia',
'South Africa',
'South Korea',
'Spain',
'Sri Lanka',
'Sweden',
'Switzerland',
'Taiwan',
'Tajikistan',
```

```
 'Tanzania',
 'Thailand',
 'The Bahamas',
 'Togo',
 'Trinidad and Tobago',
 'Turkey',
 'Uganda',
 'Ukraine',
 'United Arab Emirates',
 'United Kingdom',
 'United States',
 'Uruguay',
 'Venezuela',
 'Vietnam',
 'Yemen',
 'Zambia',
 'Zimbabwe']
```

[ ]:
```
countries = national_policy.groupby(['Country','Code'])['measure_dict'].count().
↪reset_index()
```

[ ]:
```
countries
```

[ ]:
```
                  Country Code  measure_dict
0             Afghanistan  AFG            15
1                 Albania  ALB            21
2                 Algeria  DZA            35
3                  Angola  AGO            36
4    Antigua and Barbuda  ATG            22
..                     …  …             …
188             Venezuela  VEN            25
189               Vietnam  VNM            30
190                 Yemen  YEM            17
191                Zambia  ZMB             5
192              Zimbabwe  ZWE            10

[193 rows x 3 columns]
```

[ ]:
```
country_in_mob = filtered['country_region'].unique().compute()
```

[ ]:
```
len(country_in_mob)
```

[ ]:
```
135
```

[ ]:
```
len(np.intersect1d(country_in_mob, countries['Country']))
```

[ ]:
```
130
```

[ ]:
```
country_in_both = np.intersect1d(country_in_mob, countries['Country'])
```

[ ]:
```
sorted(country_in_mob[~country_in_mob.isin(country_in_both)])
```

[ ]:
```
['Aruba', "Côte d'Ivoire", 'Puerto Rico', 'Réunion', 'Taiwan']
```

[ ]:
```
for x in sorted(countries['Country']):
    print(x)
```

```
Afghanistan
Albania
Algeria
Angola
Antigua and Barbuda
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bhutan
Bolivia
Bosnia and Herzegovina
Botswana
Brazil
Brunei
Bulgaria
Burkina Faso
Burundi
CAR
Cambodia
Cameroon
Canada
Cape Verde
Chad
Chile
China
Colombia
Comoros
Congo
Costa Rica
Croatia
```

Cuba
Cyprus
Czechia
C te d'Ivoire
DPRK
DRC
Denmark
Djibouti
Dominica
Dominican Republic
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Eswatini
Ethiopia
Fiji
Finland
France
Gabon
Gambia
Georgia
Germany
Ghana
Greece
Grenada
Guatemala
Guinea
Guinea-Bissau
Guyana
Haiti
Honduras
Hong Kong
Hungary
Iceland
India
Indonesia
Iran
Iraq
Ireland
Israel
Italy
Jamaica
Japan
Jordan
Kazakhstan

Kenya
Kiribati
Kuwait
Kyrgyzstan
Laos
Latvia
Lebanon
Lesotho
Liberia
Libya
Liechtenstein
Lithuania
Luxembourg
Madagascar
Malawi
Malaysia
Maldives
Mali
Malta
Marshall Islands
Mauritania
Mauritius
Mexico
Micronesia
Moldova
Mongolia
Montenegro
Morocco
Mozambique
Myanmar (Burma)
Namibia
Nauru
Nepal
Netherlands
New Zealand
Nicaragua
Niger
Nigeria
North Macedonia
Norway
Oman
Pakistan
Palau
Palestine
Panama
Papua New Guinea
Paraguay
Peru

Philippines
Poland
Portugal
Qatar
Romania
Russia
Rwanda
Samoa
San Marino
Sao Tome and Principe
Saudi Arabia
Senegal
Serbia
Seychelles
Sierra Leone
Singapore
Slovakia
Slovenia
Solomon Islands
Somalia
South Africa
South Korea
South Sudan
Spain
Sri Lanka
St. Kitts and Nevis
St. Lucia
St. Vincent and the Grenadines
Sudan
Suriname
Sweden
Switzerland
Syria
Tajikistan
Tanzania
Thailand
The Bahamas
Timor Leste
Togo
Tonga
Trinidad and Tobago
Tunisia
Turkey
Turkmenistan
Tuvalu
Uganda
Ukraine
United Arab Emirates

```
United Kingdom
United States
Uruguay
Uzbekistan
Vanuatu
Venezuela
Vietnam
Yemen
Zambia
Zimbabwe
```

Aruba - not in policy Czechia - Czech Republic Côte d'Ivoire - CÃ´te d'Ivoire Myanmar (Burma) - Myanmar North Macedonia - Macedonia Puerto Rico - not in policy Réunion - not in policy South Korea - Korea, Republic of Taiwan - not in policy The Bahamas - Bahamas

[ ]:
```
# Change C te d'Ivoire into to Côte d'Ivoire in national_policy
national_policy['Country'] = national_policy['Country'].replace({"C te␣
↪d'Ivoire" : "Côte d'Ivoire"})
```

[ ]:
```
countries = national_policy.groupby(['Country','Code'])['measure_dict'].count().
↪reset_index()
country_in_mob = filtered['country_region'].unique().compute()
country_in_both = np.intersect1d(country_in_mob, countries['Country'])
sorted(country_in_mob[~country_in_mob.isin(country_in_both)])
```

[ ]:
```
['Aruba', 'Puerto Rico', 'Réunion', 'Taiwan']
```

[ ]:
```
# For the country mobility filter only those that are part inc country_in_both
country_mobility = filtered[filtered['country_region'].isin(country_in_both)]
country_mobility.sample(frac=0.005, random_state=1).compute()
```

[ ]:
```
       country_region_code        country_region       date  \
13905                   AO                Angola 2021-01-12
7957                    AF           Afghanistan 2020-09-19
333                     AE  United Arab Emirates 2021-01-13
429500                  AT               Austria 2020-09-08
18117                   AR             Argentina 2020-06-24
…                      …                     …          …
493706                  YE                 Yemen 2020-05-04
506376                  ZW              Zimbabwe 2020-05-12
410222                  UY               Uruguay 2020-07-12
431494                  VN               Vietnam 2020-08-08
430406                  VE             Venezuela 2020-04-16

        retail_and_recreation_percent_change_from_baseline  \
13905                                                -6.0
7957                                                 35.0
333                                                 -18.0
```

```
429500                                                 2.0
18117                                                -62.0
…                                                    …
493706                                               -14.0
506376                                               -35.0
410222                                               -37.0
431494                                               -20.0
430406                                               -59.0

        grocery_and_pharmacy_percent_change_from_baseline  \
13905                                                  0.0
7957                                                  21.0
333                                                   -1.0
429500                                                -5.0
18117                                                -21.0
…                                                    …
493706                                                -9.0
506376                                               -26.0
410222                                               -13.0
431494                                                 2.0
430406                                               -34.0

        parks_percent_change_from_baseline  \
13905                                 28.0
7957                                  15.0
333                                  -31.0
429500                               112.0
18117                                -83.0
…                                    …
493706                               -17.0
506376                               -30.0
410222                               -66.0
431494                               -25.0
430406                               -50.0

        transit_stations_percent_change_from_baseline  \
13905                                           -3.0
7957                                           -11.0
333                                            -34.0
429500                                         -13.0
18117                                          -53.0
…                                              …
493706                                         -20.0
506376                                         -53.0
410222                                         -42.0
431494                                         -21.0
430406                                         -60.0
```

```
        workplaces_percent_change_from_baseline  \
13905                                     -9.0
7957                                       1.0
333                                      -20.0
429500                                   -28.0
18117                                    -27.0
…                                          …
493706                                   -14.0
506376                                   -31.0
410222                                    -8.0
431494                                    -8.0
430406                                   -53.0

        residential_percent_change_from_baseline
13905                                        4.0
7957                                         1.0
333                                          7.0
429500                                       1.0
18117                                       17.0
…                                            …
493706                                       5.0
506376                                      29.0
410222                                      11.0
431494                                      10.0
430406                                      22.0

[230 rows x 9 columns]
```

3.4 Save the data

[ ]:
```
country_stat.to_csv('data/cleaned/Country_stat.csv')
national_policy.to_csv('data/cleaned/National_policy.csv')
```

3.5 Merge the mobility data with the policy data

[ ]:
```
national_policy
```

[ ]:
```
     Code      Country implementation_date  \
0     AFG  Afghanistan          2020-01-26
1     AFG  Afghanistan          2020-01-27
2     AFG  Afghanistan          2020-02-01
3     AFG  Afghanistan          2020-02-02
4     AFG  Afghanistan          2020-02-12
…    …           …                  …
4942  ZWE     Zimbabwe          2020-05-04
4943  ZWE     Zimbabwe          2020-05-05
```

```
4944  ZWE     Zimbabwe          2020-05-17
4945  ZWE     Zimbabwe          2020-05-21
4946  ZWE     Zimbabwe          2020-10-01

                                          measure_dict  \
0     {'national_induction': ['Health screenings in …
1     {'national_induction': ['Health screenings in …
2             {'national_induction': ['Border checks']}
3     {'national_induction': ['Isolation and quarant…
4     {'local_induction': ['Isolation and quarantine…
…                                                   …
4942      {'national_induction': ['Partial lockdown']}
4943  {'national_phaseout': ['Limit public gathering…
4944  {'national_induction': ['Partial lockdown'], '…
4945  {'national_phaseout': ['Limit public gathering…
4946  {'national_induction': ['Health screenings in …

      national_induction_num  national_phaseout_num  local_induction_num  \
0                          1                      0                    0
1                          2                      0                    0
2                          1                      0                    0
3                          1                      0                    0
4                          1                      0                    1
…                        …                      …                    …
4942                       1                      0                    0
4943                       0                      0                    0
4944                       1                      0                    0
4945                       0                      0                    0
4946                       1                      0                    0

      local_phaseout_num
0                      0
1                      0
2                      0
3                      0
4                      0
…                    …
4942                   0
4943                   0
4944                   0
4945                   0
4946                   0

[4947 rows x 8 columns]
```

[ ]:
```
countries
```

[ ]:
```
                  Country Code  measure_dict
0             Afghanistan  AFG            15
1                 Albania  ALB            21
2                 Algeria  DZA            35
3                  Angola  AGO            36
4    Antigua and Barbuda  ATG            22
..                     …  …             …
188             Venezuela  VEN            25
189               Vietnam  VNM            30
190                 Yemen  YEM            17
191                Zambia  ZMB             5
192              Zimbabwe  ZWE            10

[193 rows x 3 columns]
```

[ ]:
```
# rename country_mobility column 'country_region' to 'Country'
country_mobility = country_mobility.rename(columns = {'country_region':
 ↪'Country'})
# rename national_policy column 'implementation_date' to 'date'
national_policy = national_policy.rename(columns = {'implementation_date':
 ↪'date'})
# merge country_mobility with Country and Code from national_policy
countries = national_policy.groupby(['Country','Code'])['measure_dict'].count().
 ↪reset_index()
country_mobility = country_mobility.merge(countries[['Country', 'Code']], on=␣
 ↪['Country'], how = 'left')
# drop country in national_policy
national_policy = national_policy.drop(columns='Country')
# merge the two
timeline = country_mobility.merge(national_policy, on=['Code', 'date'], how =␣
 ↪'left')
# drop country_region_code
timeline = timeline.drop(columns='country_region_code')
timeline.sample(frac=0.005, random_state=1).compute()
```

[ ]:
```
                    Country       date  \
1337                 Angola 2021-01-12
524             Afghanistan 2020-09-19
333   United Arab Emirates 2021-01-13
1915                Austria 2020-09-08
1487              Argentina 2020-06-24
…                     …        …
1135                  Yemen 2020-05-04
2199               Zimbabwe 2020-05-12
148                Uruguay 2020-07-12
879                Vietnam 2020-08-08
413              Venezuela 2020-04-16
```

```
      retail_and_recreation_percent_change_from_baseline  \
1337                                               -6.0
524                                                35.0
333                                               -18.0
1915                                                2.0
1487                                              -62.0
…                                                 …
1135                                              -14.0
2199                                              -35.0
148                                               -37.0
879                                               -20.0
413                                               -59.0

      grocery_and_pharmacy_percent_change_from_baseline  \
1337                                                0.0
524                                                21.0
333                                                -1.0
1915                                               -5.0
1487                                              -21.0
…                                                 …
1135                                               -9.0
2199                                              -26.0
148                                               -13.0
879                                                 2.0
413                                               -34.0

      parks_percent_change_from_baseline  \
1337                                28.0
524                                 15.0
333                                -31.0
1915                               112.0
1487                               -83.0
…                                  …
1135                               -17.0
2199                               -30.0
148                                -66.0
879                                -25.0
413                                -50.0

      transit_stations_percent_change_from_baseline  \
1337                                            -3.0
524                                            -11.0
333                                            -34.0
1915                                           -13.0
1487                                           -53.0
…                                              …
```

```
1135                                          -20.0
2199                                          -53.0
148                                           -42.0
879                                           -21.0
413                                           -60.0

      workplaces_percent_change_from_baseline  \
1337                                    -9.0
524                                      1.0
333                                    -20.0
1915                                   -28.0
1487                                   -27.0
…                                      …
1135                                   -14.0
2199                                   -31.0
148                                     -8.0
879                                     -8.0
413                                    -53.0

      residential_percent_change_from_baseline Code measure_dict  \
1337                                      4.0  AGO         <NA>
524                                       1.0  AFG         <NA>
333                                       7.0  ARE         <NA>
1915                                      1.0  AUT         <NA>
1487                                     17.0  ARG         <NA>
…                                        …  …           …
1135                                      5.0  YEM         <NA>
2199                                     29.0  ZWE         <NA>
148                                      11.0  URY         <NA>
879                                      10.0  VNM         <NA>
413                                      22.0  VEN         <NA>

      national_induction_num  national_phaseout_num  local_induction_num  \
1337                     NaN                    NaN                  NaN
524                      NaN                    NaN                  NaN
333                      NaN                    NaN                  NaN
1915                     NaN                    NaN                  NaN
1487                     NaN                    NaN                  NaN
…                        …                      …                    …
1135                     NaN                    NaN                  NaN
2199                     NaN                    NaN                  NaN
148                      NaN                    NaN                  NaN
879                      NaN                    NaN                  NaN
413                      NaN                    NaN                  NaN

      local_phaseout_num
1337                NaN
```

```
524                  NaN
333                  NaN
1915                 NaN
1487                 NaN
…                    …
1135                 NaN
2199                 NaN
148                  NaN
879                  NaN
413                  NaN

[230 rows x 14 columns]
```

[ ]:
```
philippines = timeline[timeline['Code']=='PHL'].compute()
philippines
```

[ ]:
```
         Country       date  \
0    Philippines 2020-02-15
1    Philippines 2020-02-16
2    Philippines 2020-02-17
3    Philippines 2020-02-18
4    Philippines 2020-02-19
..           …        …
347  Philippines 2021-01-27
348  Philippines 2021-01-28
349  Philippines 2021-01-29
350  Philippines 2021-01-30
351  Philippines 2021-01-31

     retail_and_recreation_percent_change_from_baseline  \
0                                                   7.0
1                                                   3.0
2                                                  -1.0
3                                                   0.0
4                                                  -3.0
..                                                  …
347                                               -40.0
348                                               -39.0
349                                               -36.0
350                                               -33.0
351                                               -35.0

     grocery_and_pharmacy_percent_change_from_baseline  \
0                                                   4.0
1                                                   5.0
2                                                   0.0
3                                                   1.0
```

```
4                                                    -3.0
..                                                    …
347                                                 -16.0
348                                                 -13.0
349                                                  -7.0
350                                                  -1.0
351                                                  -1.0

     parks_percent_change_from_baseline  \
0                                  -1.0
1                                  -2.0
2                                  -3.0
3                                  -3.0
4                                  -4.0
..                                  …
347                               -26.0
348                               -26.0
349                               -28.0
350                               -24.0
351                               -17.0

     transit_stations_percent_change_from_baseline  \
0                                              1.0
1                                             -3.0
2                                             -2.0
3                                              0.0
4                                             -6.0
..                                             …
347                                          -51.0
348                                          -50.0
349                                          -50.0
350                                          -46.0
351                                          -44.0

     workplaces_percent_change_from_baseline  \
0                                        6.0
1                                        1.0
2                                        8.0
3                                        5.0
4                                        5.0
..                                       …
347                                    -36.0
348                                    -36.0
349                                    -32.0
350                                    -18.0
351                                     -9.0
```

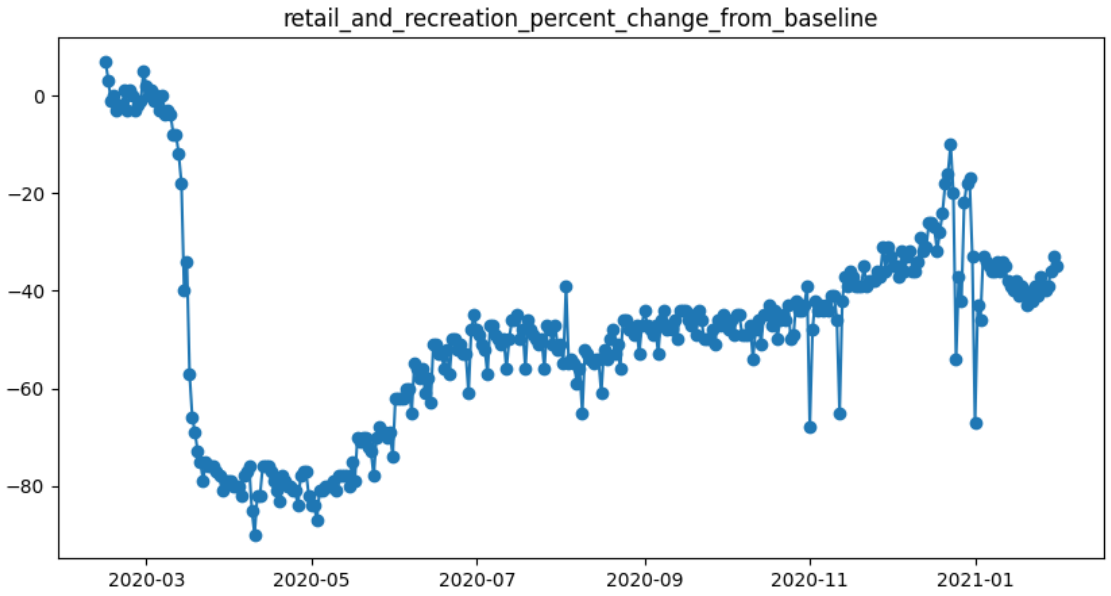
```
     residential_percent_change_from_baseline Code measure_dict  \
0                                         0.0  PHL         <NA>
1                                         1.0  PHL         <NA>
2                                         1.0  PHL         <NA>
3                                         0.0  PHL         <NA>
4                                         1.0  PHL         <NA>
..                                        …  …           …
347                                      19.0  PHL         <NA>
348                                      19.0  PHL         <NA>
349                                      19.0  PHL         <NA>
350                                      13.0  PHL         <NA>
351                                      10.0  PHL         <NA>

     national_induction_num  national_phaseout_num  local_induction_num  \
0                       NaN                    NaN                  NaN
1                       NaN                    NaN                  NaN
2                       NaN                    NaN                  NaN
3                       NaN                    NaN                  NaN
4                       NaN                    NaN                  NaN
..                      …                      …                    …
347                     NaN                    NaN                  NaN
348                     NaN                    NaN                  NaN
349                     NaN                    NaN                  NaN
350                     NaN                    NaN                  NaN
351                     NaN                    NaN                  NaN

     local_phaseout_num
0                   NaN
1                   NaN
2                   NaN
3                   NaN
4                   NaN
..                  …
347                 NaN
348                 NaN
349                 NaN
350                 NaN
351                 NaN

[352 rows x 14 columns]
```

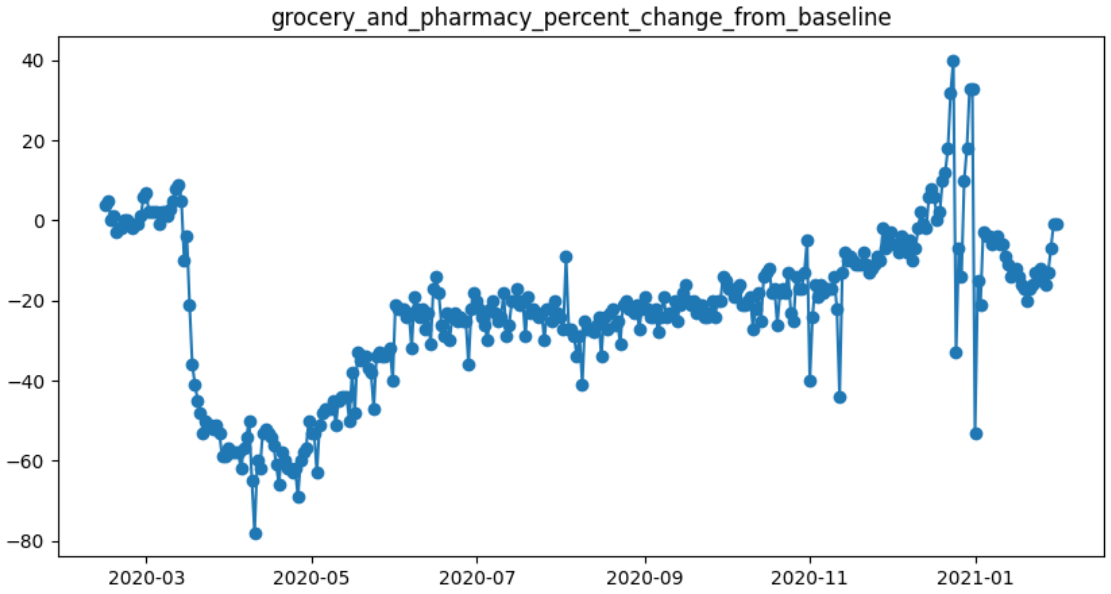
[ ]:
```
philippines.to_csv('data/cleaned/philippines.csv')
```

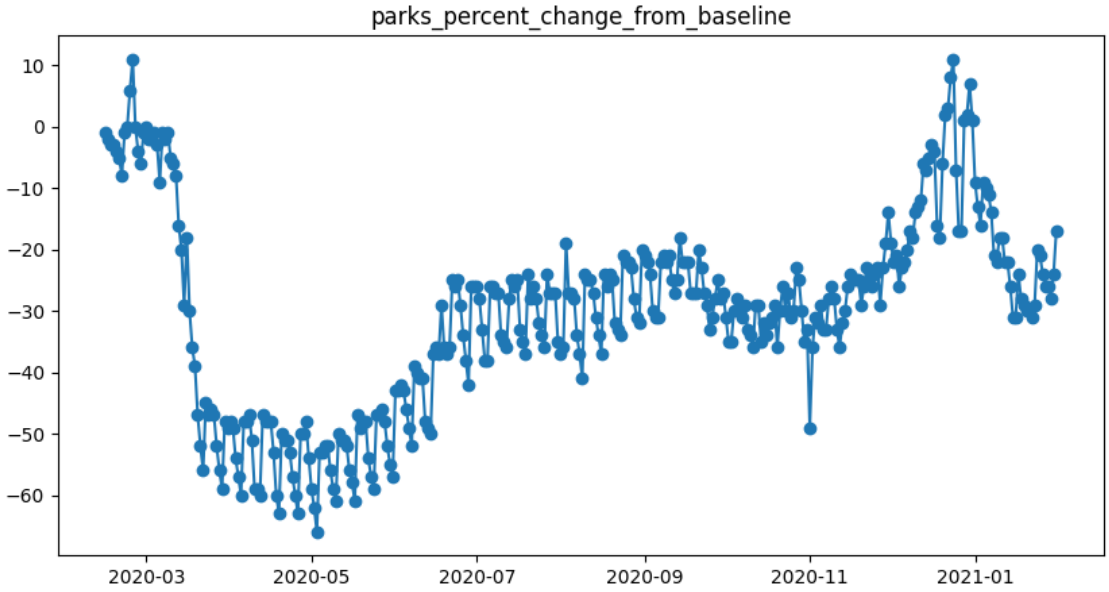
[ ]:
```
numerical = ['retail_and_recreation_percent_change_from_baseline',
             'grocery_and_pharmacy_percent_change_from_baseline',
             'parks_percent_change_from_baseline',
             'transit_stations_percent_change_from_baseline',
```

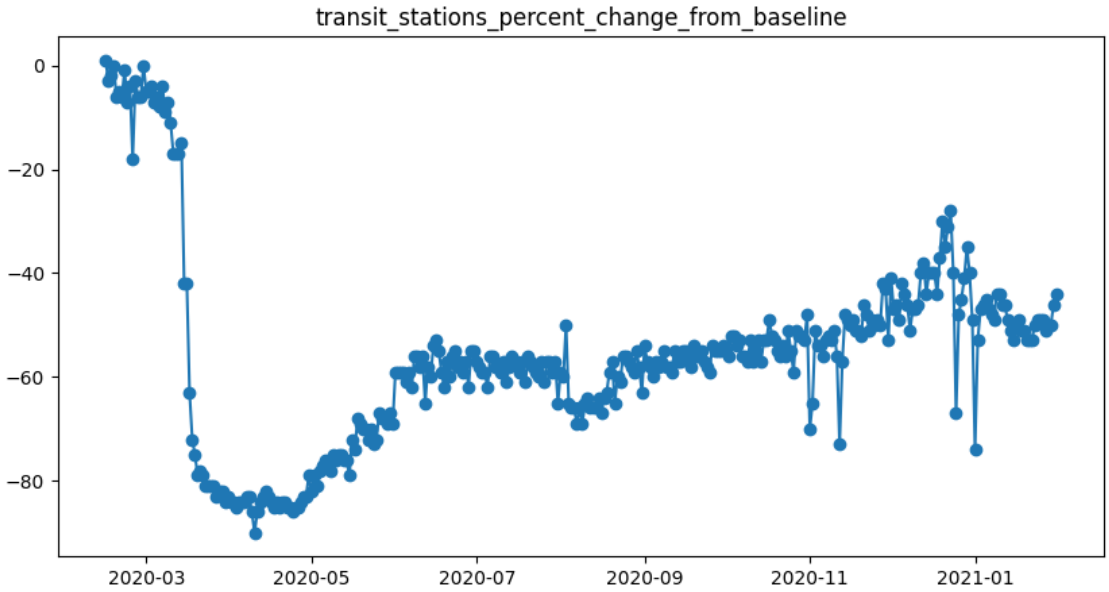
```
            'workplaces_percent_change_from_baseline',
            'residential_percent_change_from_baseline']
```

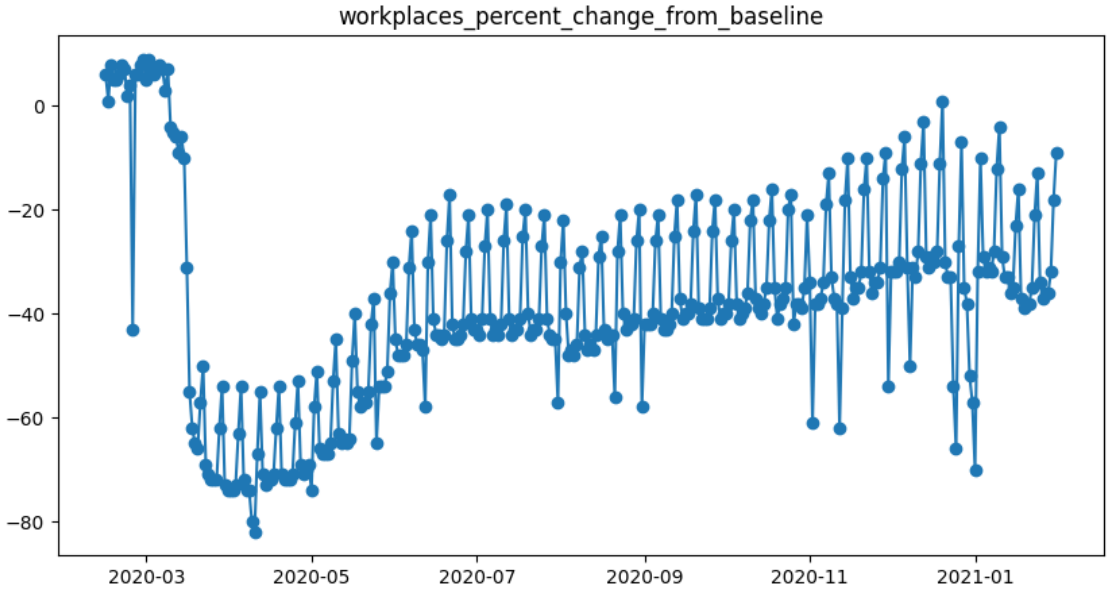
[ ]:
```
for num in numerical:
    plt.figure(figsize=(10, 5))
    plt.plot(philippines['date'], philippines[num], marker='o', linestyle='-')
    plt.title(num)
```

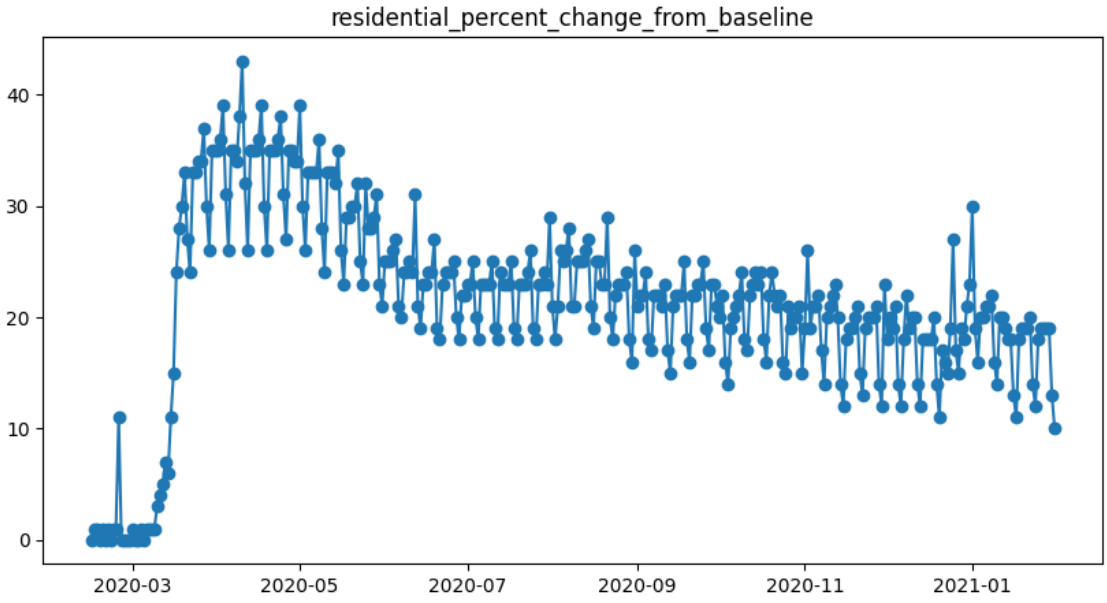





parks_percent_change_from_baseline
10
0
−10
−20
−30
−40
−50
−60
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01


transit_stations_percent_change_from_baseline
0
−20
−40
−60
−80
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01





[ ]:
```
plt.figure(figsize=(10, 5))
plt.plot(philippines['date'], philippines[num], marker='o', linestyle='-')
plt.title(num)
```