



## Data Preparation

by: Joanne Kim, Michael La, Joshua Harasaki, Tyler Chia

```
!pip install covidcast
```

```
Collecting covidcast
  Downloading covidcast-0.1.5-py3-none-any.whl (12.3 MB)
    |████████████████████| 12.3 MB 8.3 MB/s
Requirement already satisfied: matplotlib in /shared-libs/python3.7/py/lib/python3.7/site-packages (from covidcast) (3.4.3)
Collecting geopandas
  Downloading geopandas-0.10.2-py2.py3-none-any.whl (1.0 MB)
    |████████████████████| 1.0 MB 19.2 MB/s
Requirement already satisfied: numpy in /shared-libs/python3.7/py/lib/python3.7/site-packages (from covidcast) (1.19.5)
Collecting descartes
  Downloading descartes-1.1.0-py3-none-any.whl (5.8 kB)
Requirement already satisfied: requests in /shared-libs/python3.7/py/lib/python3.7/site-packages (from covidcast) (2.26.0)
Collecting imageio-ffmpeg
  Downloading imageio_ffmpeg-0.4.5-py3-none-manylinux2010_x86_64.whl (26.9 MB)
    |████████████████████| 26.9 MB 29.3 MB/s
Requirement already satisfied: pandas in /shared-libs/python3.7/py/lib/python3.7/site-packages (from covidcast) (1.2.5)
Collecting imageio
  Downloading imageio-2.13.0-py3-none-any.whl (3.3 MB)
    |████████████████████| 3.3 MB 9.3 MB/s
Collecting delphi-epidata>=0.0.11
  Downloading delphi_epidata-0.3.1-py3-none-any.whl (6.8 kB)
Requirement already satisfied: tqdm in /shared-libs/python3.7/py/lib/python3.7/site-packages (from covidcast) (4.62.3)
Collecting epiweeks
  Downloading epiweeks-2.1.3-py3-none-any.whl (5.9 kB)
Requirement already satisfied: pyparsing>=2.2.1 in /shared-libs/python3.7/py-core/lib/python3.7/site-packages (from matplotlib->covidcast) (2.4.7)
Requirement already satisfied: cycler>=0.10 in /shared-libs/python3.7/py/lib/python3.7/site-packages (from matplotlib->covidcast) (0.11.0)
Requirement already satisfied: pillow>=6.2.0 in /shared-libs/python3.7/py/lib/python3.7/site-packages (from matplotlib->covidcast) (8.4.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /shared-libs/python3.7/py/lib/python3.7/site-packages (from matplotlib->covidcast) (1.3.2)
Requirement already satisfied: python-dateutil>=2.7 in /shared-libs/python3.7/py-core/lib/python3.7/site-packages (from matplotlib->covidcast) (2.8.2)
Collecting pyproj>=2.2.0
  Downloading pyproj-3.2.1-cp37-cp37m-manylinux2010_x86_64.whl (6.3 MB)
    |████████████████████| 6.3 MB 24.0 MB/s
```

```
from datetime import date
import covidcast
import pandas as pd
import numpy as np
```

```
ca_counties = covidcast.fips_to_name("^06.*", ties_method="all")
ca_counties = list(ca_counties[0].values())
```

```
counties_string = []
for i in ca_counties:
    string = ""
    for element in i:
        string += element
    counties_string.append(string)

counties_string = counties_string[1:59] # removing 'california'
```

```
ca_counties_fips = covidcast.name_to_fips(counties_string)
ca_counties_fips
```

```
/root/venv/lib/python3.7/site-packages/covidcast/geography.py:314: UserWarning: Some inputs were not uniquely matched; returning only the first match in each case. To return all match
warnings.warn("Some inputs were not uniquely matched; returning only the first match ")
```

```
['06001',
 '06003',
 '06005',
 '06007',
 '06009',
 '06011',
 '06013',
 '06015',
 '06017',
 '06019',
 '06021',
 '06023',
 '06025',
```

```
data = covidcast.signal("indicator-combination", "confirmed_incidence_num",
geo_values= google_sum_fips)
data.head()
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value
06001	confirmed_incidence_num	2020-02-20T00:00:00.000000	2020-07-10T00:00:00.000000	141	0	5	5	0
06013	confirmed_incidence_num	2020-02-20T00:00:00.000000	2020-07-10T00:00:00.000000	141	0	5	5	0
06019	confirmed_incidence_num	2020-02-20T00:00:00.000000	2020-07-10T00:00:00.000000	141	0	5	5	0
06029	confirmed_incidence_num	2020-02-20T00:00:00.000000	2020-07-10T00:00:00.000000	141	0	5	5	0
06037	confirmed_incidence_num	2020-02-20T00:00:00.000000	2020-07-10T00:00:00.000000	141	0	5	5	0

5 rows x 9 columns

```
data.tail()
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value
06107	confirmed_incidence_num	2021-11-12T00:00:00.000000	2021-11-15T00:00:00.000000	3	0	5	5	0
06109	confirmed_incidence_num	2021-11-12T00:00:00.000000	2021-11-15T00:00:00.000000	3	0	5	5	0
06111	confirmed_incidence_num	2021-11-12T00:00:00.000000	2021-11-15T00:00:00.000000	3	0	5	5	0
06113	confirmed_incidence_num	2021-11-12T00:00:00.000000	2021-11-15T00:00:00.000000	3	0	5	5	0
06115	confirmed_incidence_num	2021-11-12T00:00:00.000000	2021-11-15T00:00:00.000000	3	0	5	5	0

5 rows x 9 columns

```
labels = data['value']
```

```
# number of observations
labels.size
```

9480

```
# looking for NA values for value column
data.isna().sum()
```

```
geo_value      0
signal         0
time_value     0
issue          0
lag            0
missing_value  0
missing_stderr 0
missing_sample_size 0
value          0
stderr        9480
sample_size    9480
geo_type       0
data_source    0
dtype: int64
```

features we are using:

smoothed\_outpatient\_cli source name : chng

smoothed\_cli source: doctor-visits

smoothed\_covid19\_from\_claims source: hospital-admissions

sum\_anosmia\_ageusia\_raw\_search, ageusia\_raw\_search, anosmia\_raw\_search source: google-symptoms

```
# many missing dates in month of november
chng = covidcast.signal("chng", "smoothed_outpatient_cli",
geo_values=google_sum_fips, start_day=date(2020, 2, 20), end_day=date(2021, 11, 12))
chng.head()
```

```
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211003 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211004 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211005 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211006 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211007 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211008 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211009 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211010 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211011 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211012 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211013 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211014 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211015 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211016 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211017 for geography 'county'
NoDataWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:425: NoDataWarning: No chng smoothed_outpatient_cli data found on 20211018 for geography 'county'
NoDataWarning)
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value
06001	smoothed_outpatient_cli	2020-02-20T00:00:00.000000	2021-02-21T00:00:00.000000	367	0	5	5	0.0226
06013	smoothed_outpatient_cli	2020-02-20T00:00:00.000000	2021-02-21T00:00:00.000000	367	0	5	5	0.0203
06019	smoothed_outpatient_cli	2020-02-20T00:00:00.000000	2021-02-21T00:00:00.000000	367	0	5	5	0.0376
06029	smoothed_outpatient_cli	2020-02-20T00:00:00.000000	2021-02-21T00:00:00.000000	367	0	5	5	0.0101
06037	smoothed_outpatient_cli	2020-02-20T00:00:00.000000	2021-02-21T00:00:00.000000	367	0	5	5	0.0041

5 rows x 9 columns

chng.tail()

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value
06075	smoothed_outpatient_cli	2021-10-02T00:00:00.000000	2021-10-07T00:00:00.000000	5	0	5	5	0.676
06077	smoothed_outpatient_cli	2021-10-02T00:00:00.000000	2021-10-07T00:00:00.000000	5	0	5	5	3.497
06081	smoothed_outpatient_cli	2021-10-02T00:00:00.000000	2021-10-07T00:00:00.000000	5	0	5	5	1.626
06085	smoothed_outpatient_cli	2021-10-02T00:00:00.000000	2021-10-07T00:00:00.000000	5	0	5	5	4.895
06111	smoothed_outpatient_cli	2021-10-02T00:00:00.000000	2021-10-07T00:00:00.000000	5	0	5	5	1.675

5 rows x 9 columns

chng.shape

(8865, 13)

```
chng.isna().sum()
```

```
geo_value      0
signal         0
time_value     0
issue         0
lag           0
missing_value  0
missing_stderr 0
missing_sample_size 0
value         0
stderr       8865
sample_size  8865
geo_type     0
data_source  0
dtype: int64
```

```
hosp = covidcast.signal("hospital-admissions", "smoothed_covid19_from_claims",
geo_values=google_sum_fips, start_day=date(2020, 2, 20), end_day=date(2021, 11, 12))
hosp.head()
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	val
06001	smoothed_covid19_from_claims	2020-02-20T00:00:00.000000	2020-06-21T00:00:00.000000	122	0	5	5	0.1
06013	smoothed_covid19_from_claims	2020-02-20T00:00:00.000000	2020-06-21T00:00:00.000000	122	0	5	5	0.1
06019	smoothed_covid19_from_claims	2020-02-20T00:00:00.000000	2020-06-21T00:00:00.000000	122	0	5	5	0.3
06037	smoothed_covid19_from_claims	2020-02-20T00:00:00.000000	2020-06-21T00:00:00.000000	122	0	5	5	0.1
06059	smoothed_covid19_from_claims	2020-02-20T00:00:00.000000	2020-06-21T00:00:00.000000	122	0	5	5	0.0

5 rows x 9 columns

```
hosp.tail()
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	val
06073	smoothed_covid19_from_claims	2021-11-12T00:00:00.000000	2021-11-28T00:00:00.000000	16	0	5	5	2.0
06075	smoothed_covid19_from_claims	2021-11-12T00:00:00.000000	2021-11-28T00:00:00.000000	16	0	5	5	1.0
06077	smoothed_covid19_from_claims	2021-11-12T00:00:00.000000	2021-11-26T00:00:00.000000	14	0	5	5	1.0
06081	smoothed_covid19_from_claims	2021-11-12T00:00:00.000000	2021-11-28T00:00:00.000000	16	0	5	5	0.0
06085	smoothed_covid19_from_claims	2021-11-12T00:00:00.000000	2021-11-28T00:00:00.000000	16	0	5	5	0.0

5 rows x 9 columns

```
hosp.shape
```

```
(9117, 13)
```

```
google_sum = covidcast.signal("google-symptoms", "sum_anosmia_ageusia_raw_search",
geo_values=google_sum_fips, start_day=date(2020, 2, 20), end_day=date(2021, 11, 12))
google_sum.head()
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	val
06001	sum_anosmia_ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.1
06013	sum_anosmia_ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.1
06037	sum_anosmia_ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.0
06059	sum_anosmia_ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.1
06065	sum_anosmia_ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.1

```
google_ageusia = covidcast.signal("google-symptoms", "ageusia_raw_search",
geo_values=google_sum_fips, start_day=date(2020, 2, 20), end_day=date(2021, 11, 12))
google_ageusia.head()
```

	geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value float64
0	06001	ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.06
1	06037	ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.03
2	06059	ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.08
3	06065	ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.04
4	06073	ageusia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.04

5 rows x 10 columns

```
google_anosmia = covidcast.signal("google-symptoms", "anosmia_raw_search",
geo_values=google_sum_fips, start_day=date(2020, 2, 20), end_day=date(2021, 11, 12))
google_anosmia.head()
```

```
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210528 for geography 'county': error:
RuntimeWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210530 for geography 'county': error:
RuntimeWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210531 for geography 'county': error:
RuntimeWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210607 for geography 'county': error:
RuntimeWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210614 for geography 'county': error:
RuntimeWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210616 for geography 'county': error:
RuntimeWarning)
/root/venv/lib/python3.7/site-packages/covidcast/covidcast.py:429: RuntimeWarning: Problem obtaining google-symptoms anosmia_raw_search data on 20210618 for geography 'county': error:
RuntimeWarning)
```

	geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value float64
0	06001	anosmia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.06
1	06013	anosmia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.14
2	06037	anosmia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.06
3	06059	anosmia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.09
4	06065	anosmia_raw_search	2020-02-20T00:00:00.000000	2021-01-14T00:00:00.000000	329	0	5	5	0.06

5 rows x 10 columns

```
# checking to see if counties match up in different signals
fips = []
for i in ageusia_fips:
    if i in hosp_fips:
        fips.append(i)
```

```
doc = covidcast.signal("doctor-visits", "smoothed_cli",
geo_values=google_sum_fips, start_day=date(2020, 2, 20), end_day=date(2021, 11, 12))
doc.head()
```

	geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value float64	st
0	06001	smoothed_cli	2020-02-20T00:00:00.000000	2020-06-09T00:00:00.000000	110	0	5	5	0	No
1	06013	smoothed_cli	2020-02-20T00:00:00.000000	2020-06-09T00:00:00.000000	110	0	5	5	0	No
2	06019	smoothed_cli	2020-02-20T00:00:00.000000	2020-06-09T00:00:00.000000	110	0	5	5	0.051447	No
3	06029	smoothed_cli	2020-02-20T00:00:00.000000	2020-06-09T00:00:00.000000	110	0	5	5	0	No
4	06037	smoothed_cli	2020-02-20T00:00:00.000000	2020-06-09T00:00:00.000000	110	0	5	5	0.13478	No

```
doc.tail()
```

geo_value object	signal object	time_value datetime64[ns]	issue datetime64[ns]	lag int64	missing_value int64	missing_stderr int64	missing_sample_size int64	value float64	std float64
06103	smoothed_cli	2021-11-12T00:00:00.000000	2021-11-27T00:00:00.000000	15	0	5	5	2.29704	NaN
06107	smoothed_cli	2021-11-12T00:00:00.000000	2021-11-27T00:00:00.000000	15	0	5	5	3.001187	NaN
06111	smoothed_cli	2021-11-12T00:00:00.000000	2021-11-27T00:00:00.000000	15	0	5	5	1.986365	NaN
06113	smoothed_cli	2021-11-12T00:00:00.000000	2021-11-27T00:00:00.000000	15	0	5	5	0.976299	NaN
06115	smoothed_cli	2021-11-12T00:00:00.000000	2021-11-27T00:00:00.000000	15	0	5	5	1.265493	NaN

5 rows x 10 columns

```
np.array(doc).shape
```

(9480, 13)

```
doc.isna().sum()
```

```

geo_value      0
signal         0
time_value     0
issue          0
lag            0
missing_value  0
missing_stderr 0
missing_sample_size 0
value          0
stderr        28440
sample_size    28440
geo_type       0
data_source    0
dtype: int64
    
```

```
merged = covidcast.aggregate_signals([hosp, chng, doc, google_sum, google_anosmia, google_ageusia, data])
```



Expand rows 5 - 9474

06075	2021-11-12T00:00:00.000000	2021-11-28 00:00:00	16	0	5	5
06077	2021-11-12T00:00:00.000000	2021-11-26 00:00:00	14	0	5	5
06081	2021-11-12T00:00:00.000000	2021-11-28 00:00:00	16	0	5	5
06085	2021-11-12T00:00:00.000000	2021-11-28 00:00:00	16	0	5	5
06111	2021-11-12T00:00:00.000000	NaT	nan	nan	nan	nan

9480 rows x 7 columns

merged.columns

```
Index(['geo_value', 'time_value',
      'hospital-admissions_smoothed_covid19_from_claims_0_issue',
      'hospital-admissions_smoothed_covid19_from_claims_0_lag',
      'hospital-admissions_smoothed_covid19_from_claims_0_missing_value',
      'hospital-admissions_smoothed_covid19_from_claims_0_missing_stderr',
      'hospital-admissions_smoothed_covid19_from_claims_0_missing_sample_size',
      'hospital-admissions_smoothed_covid19_from_claims_0_value',
      'hospital-admissions_smoothed_covid19_from_claims_0_stderr',
      'hospital-admissions_smoothed_covid19_from_claims_0_sample_size',
      'chng_smoothed_outpatient_cli_1_issue',
      'chng_smoothed_outpatient_cli_1_lag',
      'chng_smoothed_outpatient_cli_1_missing_value',
      'chng_smoothed_outpatient_cli_1_missing_stderr',
      'chng_smoothed_outpatient_cli_1_missing_sample_size',
      'chng_smoothed_outpatient_cli_1_value',
      'chng_smoothed_outpatient_cli_1_stderr',
      'chng_smoothed_outpatient_cli_1_sample_size',
      'doctor-visits_smoothed_cli_2_issue',
      'doctor-visits_smoothed_cli_2_lag',
      'doctor-visits_smoothed_cli_2_missing_value',
      'doctor-visits_smoothed_cli_2_missing_stderr',
      'doctor-visits_smoothed_cli_2_missing_sample_size',
      'doctor-visits_smoothed_cli_2_value',
      'doctor-visits_smoothed_cli_2_stderr',
      'doctor-visits_smoothed_cli_2_sample_size',
      'google-symptoms_sum_anosmia_ageusia_raw_search_3_issue',
      'google-symptoms_sum_anosmia_ageusia_raw_search_3_lag',
      'google-symptoms_sum_anosmia_ageusia_raw_search_3_missing_stderr',
      'google-symptoms_sum_anosmia_ageusia_raw_search_3_missing_sample_size',
      'google-symptoms_sum_anosmia_ageusia_raw_search_3_value',
      'google-symptoms_sum_anosmia_ageusia_raw_search_3_stderr',
```

```
x = merged.drop(columns = [
    'hospital-admissions_smoothed_covid19_from_claims_0_issue',
    'hospital-admissions_smoothed_covid19_from_claims_0_lag',
    'hospital-admissions_smoothed_covid19_from_claims_0_missing_value',
    'hospital-admissions_smoothed_covid19_from_claims_0_missing_stderr',
    'hospital-admissions_smoothed_covid19_from_claims_0_missing_sample_size',
    'hospital-admissions_smoothed_covid19_from_claims_0_stderr',
    'hospital-admissions_smoothed_covid19_from_claims_0_sample_size',
    'chng_smoothed_outpatient_cli_1_issue',
    'chng_smoothed_outpatient_cli_1_lag',
    'chng_smoothed_outpatient_cli_1_missing_value',
    'chng_smoothed_outpatient_cli_1_missing_stderr',
    'chng_smoothed_outpatient_cli_1_missing_sample_size',
    'chng_smoothed_outpatient_cli_1_stderr',
    'chng_smoothed_outpatient_cli_1_sample_size',
    'doctor-visits_smoothed_cli_2_issue',
    'doctor-visits_smoothed_cli_2_lag',
    'doctor-visits_smoothed_cli_2_missing_value',
    'doctor-visits_smoothed_cli_2_missing_stderr',
    'doctor-visits_smoothed_cli_2_missing_sample_size',
    'doctor-visits_smoothed_cli_2_stderr',
    'doctor-visits_smoothed_cli_2_sample_size',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_issue',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_lag',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_missing_value',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_missing_stderr',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_missing_sample_size',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_stderr',
    'google-symptoms_sum_anosmia_ageusia_raw_search_3_sample_size',
    'google-symptoms_anosmia_raw_search_4_issue',
    'google-symptoms_anosmia_raw_search_4_lag',
    'google-symptoms_anosmia_raw_search_4_missing_value',
    'google-symptoms_anosmia_raw_search_4_missing_stderr',
    'google-symptoms_anosmia_raw_search_4_missing_sample_size',
    'google-symptoms_anosmia_raw_search_4_stderr',
    'google-symptoms_anosmia_raw_search_4_sample_size',
    'google-symptoms_ageusia_raw_search_5_issue',
    'google-symptoms_ageusia_raw_search_5_lag',
    'google-symptoms_ageusia_raw_search_5_missing_value',
    'google-symptoms_ageusia_raw_search_5_missing_stderr',
    'google-symptoms_ageusia_raw_search_5_missing_sample_size',
    'google-symptoms_ageusia_raw_search_5_stderr',
    'google-symptoms_ageusia_raw_search_5_sample_size',
    'indicator-combination_confirmed_incidence_num_6_issue',
    'indicator-combination_confirmed_incidence_num_6_lag',
    'indicator-combination_confirmed_incidence_num_6_missing_value',
    'indicator-combination_confirmed_incidence_num_6_missing_stderr',
    'indicator-combination_confirmed_incidence_num_6_missing_sample_size',
    'indicator-combination_confirmed_incidence_num_6_stderr',
    'indicator-combination_confirmed_incidence_num_6_sample_size',
    'geo_type'])
```

x

geo_value	object	time_value	datetime64[ns]	hospital-admissions_s...	chng_smoothed_outpat...	doctor-visits_smooth...	google-symptoms_sum_a...	google-symptoms_anos...
06001	6.7%	2020-02-20 00:00:00 - 2021-		0.050273 - 31.394108	0.002845 - 35.921154	0.0 - 48.899837	0.04 - 3.61	0.04 - 2.86
06013	6.7%							

	13 others ... 86.7%						
0	06001	2020-02-20T00:00:00.000000	0.110992	0.022051	0	0.12	0.06
1	06013	2020-02-20T00:00:00.000000	0.118745	0.020335	0	0.14	0.14
2	06019	2020-02-20T00:00:00.000000	0.353801	0.037049	0.051447	nan	nan
3	06029	2020-02-20T00:00:00.000000	nan	0.01015	0	nan	nan
4	06037	2020-02-20T00:00:00.000000	0.112637	0.004175	0.13478	0.09	0.06

Expand rows 5 - 9474

9475	06075	2021-11-12T00:00:00.000000	1.230602	nan	2.764137	0.16	0.1
9476	06077	2021-11-12T00:00:00.000000	1.706939	nan	2.002706	0.13	0.13
9477	06081	2021-11-12T00:00:00.000000	0.185958	nan	2.497344	nan	nan
9478	06085	2021-11-12T00:00:00.000000	0.103307	nan	9.675555	0.15	0.09
9479	06111	2021-11-12T00:00:00.000000	nan	nan	1.986365	0.31	0.17

9480 rows x 9 columns

x.columns

```
Index(['geo_value', 'time_value',
       'hospital-admissions_smoothed_covid19_from_claims_0_value',
       'chn_g_smoothed_outpatient_cli_1_value',
       'doctor-visits_smoothed_cli_2_value',
       'google-symptoms_sum_anosmia_ageusia_raw_search_3_value',
       'google-symptoms_anosmia_raw_search_4_value',
       'google-symptoms_ageusia_raw_search_5_value',
       'indicator-combination_confirmed_incidence_num_6_value'],
      dtype='object')
```

x.isna().sum()

```
geo_value                0
time_value               0
hospital-admissions_smoothed_covid19_from_claims_0_value    363
chn_g_smoothed_outpatient_cli_1_value                       615
doctor-visits_smoothed_cli_2_value                         0
google-symptoms_sum_anosmia_ageusia_raw_search_3_value     710
google-symptoms_anosmia_raw_search_4_value                 986
google-symptoms_ageusia_raw_search_5_value                 2277
indicator-combination_confirmed_incidence_num_6_value      0
dtype: int64
```

```
np.unique(x[x['google-symptoms_ageusia_raw_search_5_value'].isna()]['geo_value'])
# these are the unique counties that the ageusia column in the merged dataset is na for ?
```

```
array(['06001', '06013', '06019', '06029', '06065', '06067', '06071',
       '06075', '06077', '06081', '06085', '06111'], dtype=object)
```

x.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9480 entries, 0 to 9479
Data columns (total 9 columns):
#   Column                                                                 Non-Null Count  Dtype
---  -
0   geo_value                                                            9480 non-null   object
1   time_value                                                            9480 non-null   datetime64[ns]
2   hospital-admissions_smoothed_covid19_from_claims_0_value          9117 non-null   float64
3   chng_smoothed_outpatient_cli_1_value                                8865 non-null   float64
4   doctor-visits_smoothed_cli_2_value                                  9480 non-null   float64
5   google-symptoms_sum_anosmia_ageusia_raw_search_3_value            8770 non-null   float64
6   google-symptoms_anosmia_raw_search_4_value                        8494 non-null   float64
7   google-symptoms_ageusia_raw_search_5_value                        7203 non-null   float64
8   indicator-combination_confirmed_incidence_num_6_value             9480 non-null   float64
dtypes: datetime64[ns](1), float64(7), object(1)
memory usage: 998.7+ KB
```



```
min(x['indicator-combination_confirmed_incidence_num_6_value'])
```

-12459.0

```
# seeing negative values for our labels????
x[x['indicator-combination_confirmed_incidence_num_6_value'] < 0]
```

geo_value	object	time_value	datetime64[ns]	hospital-admissions_s_ f_	chn_g_smoothed_outpat_ flo_	doctor-visits_smoott_ flo_	google-symptoms_sum_a_ f_	google-symptoms_anos_ f_
24116	06013	2020-07-30T00:00:00.000000	0.268705 - 19.245695	4.962517	8.64017	5.173582	0.31000000000000005	0.17
24223	06071	2020-07-30T00:00:00.000000	0.268705 - 19.245695	7.800456	4.949401	4.250491	0.42000000000000004	0.25
22945	06059	2020-09-03T00:00:00.000000	0.268705 - 19.245695	3.319291	2.972764	4.198172	0.21	0.12
32003	06071	2020-09-20T00:00:00.000000	0.268705 - 19.245695	2.644268	4.267572	3.396508	0.2	0.11
34057	06019	2020-10-08T00:00:00.000000	0.268705 - 19.245695	0.306217	3.237671	3.075505	0.15	0.2037520602778432

Expand rows 5 - 23

93516	06077	2021-11-04T00:00:00.000000	0.268705 - 19.245695	4.972303	5.302123934156795	2.253787	0.32694982896237174	0.2037520602778432
94225	06059	2021-11-09T00:00:00.000000	0.268705 - 19.245695	4.121999	5.302123934156795	2.247481	0.21	0.11
94226	06065	2021-11-09T00:00:00.000000	0.268705 - 19.245695	4.720285	5.302123934156795	3.63074	0.3	0.14
94300	06075	2021-11-09T00:00:00.000000	0.268705 - 19.245695	1.606974	5.302123934156795	3.625614	0.15	0.15
94445	06077	2021-11-10T00:00:00.000000	0.268705 - 19.245695	3.041163	5.302123934156795	1.439409	0.32694982896237174	0.2037520602778432

15 rows x 9 columns

```
# converting all negative labels to positive
x['indicator-combination_confirmed_incidence_num_6_value'] = abs(x['indicator-combination_confirmed_incidence_num_6_value'])
# drop ageusia since missing a county
x = x.drop(columns = "google-symptoms_ageusia_raw_search_5_value")
x
```

geo_value	object	time_value	datetime64[ns]	hospital-admissions_s_ f_	chn_g_smoothed_outpat_ flo_	doctor-visits_smoott_ flo_	google-symptoms_sum_a_ f_	google-symptoms_anos_ f_
0	06001	2020-02-20T00:00:00.000000	0.058273 - 31.394108	0.110992	0.022051	0	0.12	0.06
1	06013	2020-02-20T00:00:00.000000	0.058273 - 31.394108	0.118745	0.020335	0	0.14	0.14
2	06019	2020-02-20T00:00:00.000000	0.058273 - 31.394108	0.353001	0.037049	0.051447	nan	nan
3	06029	2020-02-20T00:00:00.000000	0.058273 - 31.394108	nan	0.01015	0	nan	nan
4	06037	2020-02-20T00:00:00.000000	0.058273 - 31.394108	0.112637	0.004175	0.13478	0.09	0.06

Expand rows 5 - 9474

9475	06075	2021-11-12T00:00:00.000000	0.058273 - 31.394108	1.230602	nan	2.764137	0.16	0.1
9476	06077	2021-11-12T00:00:00.000000	0.058273 - 31.394108	1.706939	nan	2.002706	0.13	0.13
9477	06081	2021-11-12T00:00:00.000000	0.058273 - 31.394108	0.185958	nan	2.497344	nan	nan
9478	06085	2021-11-12T00:00:00.000000	0.058273 - 31.394108	0.103307	nan	9.675555	0.15	0.09
9479	06111	2021-11-12T00:00:00.000000	0.058273 - 31.394108	nan	nan	1.986365	0.31	0.17

9480 rows x 9 columns

```
x.isna().sum()
```

```
geo_value          0
time_value         0
hospital-admissions_smoothed_covid19_from_claims_0_value 363
chng_smoothed_outpatient_cli_1_value      615
doctor-visits_smoothed_cli_2_value        0
google-symptoms_sum_anosmia_ageusia_raw_search_3_value  710
google-symptoms_anosmia_raw_search_4_value  986
indicator-combination_confirmed_incidence_num_6_value   0
dtype: int64
```

```
# imputing by forward filling based on previous observation in each county
```

```
updated_x = x
updated_x['hospital-admissions_smoothed_covid19_from_claims_0_value'] = x.groupby('geo_value')['hospital-admissions_smoothed_covid19_from_claims_0_value'].fillna(method='ffill')
updated_x['chng_smoothed_outpatient_cli_1_value'] = x.groupby('geo_value')['chng_smoothed_outpatient_cli_1_value'].fillna(method='ffill')
updated_x['google-symptoms_sum_anosmia_ageusia_raw_search_3_value'] = x.groupby('geo_value')['google-symptoms_sum_anosmia_ageusia_raw_search_3_value'].fillna(method='ffill')
updated_x['google-symptoms_anosmia_raw_search_4_value'] = x.groupby('geo_value')['google-symptoms_anosmia_raw_search_4_value'].fillna(method='ffill')
updated_x
```



	geo_value	time_value	hospital-admissions_s...	chng_smoothed_outpat...	doctor-visits_smoote...	google-symptoms_sum_a...	google-symptoms_anos...
0	06001	2020-02-20T00:00:00.000000	0.110992	0.022051	0	0.12	0.06
1	06013	2020-02-20T00:00:00.000000	0.118745	0.020335	0	0.14	0.14
2	06019	2020-02-20T00:00:00.000000	0.353801	0.037049	0.051447	nan	nan
3	06029	2020-02-20T00:00:00.000000	nan	0.01015	0	nan	nan
4	06037	2020-02-20T00:00:00.000000	0.112637	0.004175	0.13478	0.09	0.06

Expand rows 5 - 9474

9475	06075	2021-11-12T00:00:00.000000	1.230602	0.6704379	2.764137	0.16	0.1
9476	06077	2021-11-12T00:00:00.000000	1.706939	3.4975448	2.002706	0.13	0.13
9477	06081	2021-11-12T00:00:00.000000	0.185958	1.6268549	2.497344	0.15	0.15
9478	06085	2021-11-12T00:00:00.000000	0.103307	4.0956309	9.675555	0.15	0.09
9479	06111	2021-11-12T00:00:00.000000	18.908364	1.6790429	1.986365	0.31	0.17

9480 rows x 8 columns

```
updated_x.isna().sum()
```

```
geo_value          0
time_value         0
hospital-admissions_smoothed_covid19_from_claims_0_value 28
chng_smoothed_outpatient_cli_1_value      0
doctor-visits_smoothed_cli_2_value        0
google-symptoms_sum_anosmia_ageusia_raw_search_3_value  11
google-symptoms_anosmia_raw_search_4_value  29
indicator-combination_confirmed_incidence_num_6_value   0
dtype: int64
```

```
# drop remaining NA values
```

```
updated_x = updated_x.dropna(axis = 0)
updated_x.isna().sum()
```

```
geo_value          0
time_value         0
hospital-admissions_smoothed_covid19_from_claims_0_value 0
chng_smoothed_outpatient_cli_1_value      0
doctor-visits_smoothed_cli_2_value        0
google-symptoms_sum_anosmia_ageusia_raw_search_3_value  0
google-symptoms_anosmia_raw_search_4_value  0
indicator-combination_confirmed_incidence_num_6_value   0
dtype: int64
```

```
updated_x
```

	geo_value object	time_value datetime64[ns]	hospital-admissions_s_ f_	chng_smoothed_outpat_ flo_	doctor-visits_smoos_ flo_	google-symptoms_sum_a_ f_	google-symptoms_anos_ f_
	06001 ..... 6.7% 06013 ..... 6.7% 13 others ..... 86.6%	2020-02-20 00:00:00 - 2021-02-20 00:00:00	0.058273 - 31.394108	0.002845 - 35.921154	0.0 - 48.899837	0.04 - 3.61	0.04 - 2.86
0	06001	2020-02-20T00:00:00.000000	0.110992	0.022051	0	0.12	0.06
11	06013	2020-02-20T00:00:00.000000	0.118745	0.020335	0	0.14	0.14
4	06037	2020-02-20T00:00:00.000000	0.112637	0.004175	0.13478	0.09	0.06
5	06059	2020-02-20T00:00:00.000000	0.091819	0.005578	0.077821	0.16999999999999998	0.09
6	06065	2020-02-20T00:00:00.000000	0.090382	0.009983	0	0.1	0.06

Expand rows 5 - 9420

9475	06075	2021-11-12T00:00:00.000000	1.230602	0.6704379	2.764137	0.16	0.1
9476	06077	2021-11-12T00:00:00.000000	1.706939	3.4975448	2.002706	0.13	0.13
9477	06081	2021-11-12T00:00:00.000000	0.185958	1.6268549	2.497344	0.15	0.15
9478	06085	2021-11-12T00:00:00.000000	0.103307	4.8956309	9.675555	0.15	0.09
9479	06111	2021-11-12T00:00:00.000000	18.908364	1.6790429	1.986365	0.31	0.17

9436 rows x 8 columns

updated\_x.to\_csv(r'data\_preparation.csv')