COVID19_Analysis_Texas

** Created By: Dr. Meena Kusi

In [ ]:
```
### Libraries used
```

In [341]:
```
import pandas as pd
import numpy as np

# charting
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

1. Read the data from the source

In [342]:
```
df1 = pd.read_csv("./data/overall-state-smoothed.csv", index_col='date', parse_dates=True)
```

In this data set, there are 124664 patients data and 103 features

In [343]:
```
df1.head()
```

Out[343]:

date	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	smoothe
2020-11-01	ak	female	18-34	72.0	1.4757	1.4757	
2020-10-31	ak	female	18-34	60.0	0.0000	0.0000	
2020-10-30	ak	female	18-34	54.0	0.0000	0.0000	
2020-10-29	ak	female	18-34	59.0	0.0000	0.0000	
2020-10-28	ak	female	18-34	68.0	0.0000	0.0000	

5 rows × 102 columns

In [344]:
```
# df1.groupby('state_code').count()['date']
```

Here we are going to do some analysis in Texas only so let's select the texas data

In [345]:
```
df1_tx =df1[df1['state_code']=='tx']
```

In [346]:
```
df1_tx.shape
```

Out[346]: (2448, 102)

In texas only, there are 2448 patients data with 103 features

Looking at this data there are equal number of male and female patients

In [347]:
```
# To list all the columns in "df_tx"
df1_tx.columns.tolist()
```

```
Out[347]: ['state_code',
           'gender',
           'age_bucket',
           'summed_n',
           'smoothed_pct_cli',
           'smoothed_pct_ili',
           'smoothed_pct_cli_anosmia_ageusia',
           'smoothed_pct_hh_cli',
           'smoothed_pct_cmnty_cli',
           'smoothed_pct_hh_fever',
           'smoothed_pct_hh_sore_throat',
           'smoothed_pct_hh_cough',
           'smoothed_pct_hh_shortness_of_breath',
           'smoothed_pct_hh_difficulty_breathing',
           'smoothed_mean_hh_cli_ct',
           'smoothed_mean_cmnty_cli_ct',
           'smoothed_pct_self_fever',
           'smoothed_pct_self_cough',
           'smoothed_pct_self_shortness_of_breath',
           'smoothed_pct_self_difficulty_breathing',
           'smoothed_pct_self_tiredness_or_exhaustion',
           'smoothed_pct_self_nasal_congestion',
           'smoothed_pct_self_runny_nose',
           'smoothed_pct_self_muscle_joint_aches',
           'smoothed_pct_self_sore_throat',
           'smoothed_pct_self_persistent_pain_pressure_in_chest',
           'smoothed_pct_self_nausea_vomiting',
           'smoothed_pct_self_diarrhea',
           'smoothed_pct_self_anosmia_ageusia',
           'smoothed_pct_self_other',
           'smoothed_pct_self_none_of_above',
           'smoothed_pct_self_multiple_symptoms',
           'smoothed_pct_tested_and_positive',
           'smoothed_pct_tested_and_negative',
           'smoothed_pct_tested_no_result',
           'smoothed_pct_could_not_get_tested',
           'smoothed_pct_did_not_try_to_get_tested',
           'smoothed_pct_worked_outside_home',
           'smoothed_pct_avoid_contact_all_or_most_time',
           'smoothed_mean_outside_hh_contact_at_work_ct',
           'smoothed_mean_outside_hh_contact_shopping_ct',
           'smoothed_mean_outside_hh_contact_in_social_gatherings_ct',
           'smoothed_pct_contact_covid_positive',
           'smoothed_pct_diabetes',
           'smoothed_pct_cancer',
           'smoothed_pct_heart_disease',
           'smoothed_pct_high_blood_pressure',
           'smoothed_pct_asthma',
           'smoothed_pct_chronic_lung_disease',
           'smoothed_pct_kidney_disease',
           'smoothed_pct_autoimmune_disorder',
           'smoothed_pct_no_above_medical_conditions',
           'smoothed_pct_multiple_medical_conditions',
           'smoothed_pct_cli_weighted',
           'smoothed_pct_ili_weighted',
           'smoothed_pct_cli_anosmia_ageusia_weighted',
           'smoothed_pct_hh_cli_weighted',
```

```
'smoothed_pct_cmnty_cli_weighted',
'smoothed_pct_hh_fever_weighted',
'smoothed_pct_hh_sore_throat_weighted',
'smoothed_pct_hh_cough_weighted',
'smoothed_pct_hh_shortness_of_breath_weighted',
'smoothed_pct_hh_difficulty_breathing_weighted',
'smoothed_mean_hh_cli_ct_weighted',
'smoothed_mean_cmnty_cli_ct_weighted',
'smoothed_pct_self_fever_weighted',
'smoothed_pct_self_cough_weighted',
'smoothed_pct_self_shortness_of_breath_weighted',
'smoothed_pct_self_difficulty_breathing_weighted',
'smoothed_pct_self_tiredness_or_exhaustion_weighted',
'smoothed_pct_self_nasal_congestion_weighted',
'smoothed_pct_self_runny_nose_weighted',
'smoothed_pct_self_muscle_joint_aches_weighted',
'smoothed_pct_self_sore_throat_weighted',
'smoothed_pct_self_persistent_pain_pressure_in_chest_weighted',
'smoothed_pct_self_nausea_vomiting_weighted',
'smoothed_pct_self_diarrhea_weighted',
'smoothed_pct_self_anosmia_ageusia_weighted',
'smoothed_pct_self_other_weighted',
'smoothed_pct_self_none_of_above_weighted',
'smoothed_pct_self_multiple_symptoms_weighted',
'smoothed_pct_tested_and_positive_weighted',
'smoothed_pct_tested_and_negative_weighted',
'smoothed_pct_tested_no_result_weighted',
'smoothed_pct_could_not_get_tested_weighted',
'smoothed_pct_did_not_try_to_get_tested_weighted',
'smoothed_pct_worked_outside_home_weighted',
'smoothed_pct_avoid_contact_all_or_most_time_weighted',
'smoothed_mean_outside_hh_contact_at_work_ct_weighted',
'smoothed_mean_outside_hh_contact_shopping_ct_weighted',
'smoothed_mean_outside_hh_contact_in_social_gatherings_ct_weighted',
'smoothed_pct_contact_covid_positive_weighted',
'smoothed_pct_diabetes_weighted',
'smoothed_pct_cancer_weighted',
'smoothed_pct_heart_disease_weighted',
'smoothed_pct_high_blood_pressure_weighted',
'smoothed_pct_asthma_weighted',
'smoothed_pct_chronic_lung_disease_weighted',
'smoothed_pct_kidney_disease_weighted',
'smoothed_pct_autoimmune_disorder_weighted',
'smoothed_pct_no_above_medical_conditions_weighted',
'smoothed_pct_multiple_medical_conditions_weighted']
```

2. Feature selection

We are interested only in few features. Here are the features we are going to use in our analysis

In [348]:
```
# # To keep only desired columns:
# columns_required = ['summed_n',
#  'gender',
#  'age_bucket',
#  'smoothed_pct_cli_weighted',
#  'smoothed_pct_ili_weighted',
#  'smoothed_pct_cli_anosmia_ageusia_weighted',
#  'smoothed_pct_hh_cli_weighted',
#  'smoothed_pct_cmnty_cli_weighted',
#  'smoothed_pct_hh_fever_weighted',
#  'smoothed_pct_hh_sore_throat_weighted',
#  'smoothed_pct_hh_cough_weighted',
#  'smoothed_pct_hh_shortness_of_breath_weighted',
#  'smoothed_pct_hh_difficulty_breathing_weighted',
#  'smoothed_mean_hh_cli_ct_weighted',
#  'smoothed_mean_cmnty_cli_ct_weighted',
#  'smoothed_pct_self_fever_weighted',
#  'smoothed_pct_self_cough_weighted',
#  'smoothed_pct_self_shortness_of_breath_weighted',
#  'smoothed_pct_self_difficulty_breathing_weighted',
#  'smoothed_pct_self_tiredness_or_exhaustion_weighted',
#  'smoothed_pct_self_nasal_congestion_weighted',
#  'smoothed_pct_self_runny_nose_weighted',
#  'smoothed_pct_self_muscle_joint_aches_weighted',
#  'smoothed_pct_self_sore_throat_weighted',
#  'smoothed_pct_self_persistent_pain_pressure_in_chest_weighted',
#  'smoothed_pct_self_nausea_vomiting_weighted',
#  'smoothed_pct_self_diarrhea_weighted',
#  'smoothed_pct_self_anosmia_ageusia_weighted',
#  'smoothed_pct_self_other_weighted',
#  'smoothed_pct_self_none_of_above_weighted',
#  'smoothed_pct_self_multiple_symptoms_weighted',
#  'smoothed_pct_tested_and_positive_weighted',
#  'smoothed_pct_tested_and_negative_weighted',
#  'smoothed_pct_tested_no_result_weighted',
#  'smoothed_pct_could_not_get_tested_weighted',
#  'smoothed_pct_did_not_try_to_get_tested_weighted',
#  'smoothed_pct_worked_outside_home_weighted',
#  'smoothed_pct_avoid_contact_all_or_most_time_weighted',
#  'smoothed_mean_outside_hh_contact_at_work_ct_weighted',
#  'smoothed_mean_outside_hh_contact_shopping_ct_weighted',
#  'smoothed_mean_outside_hh_contact_in_social_gatherings_ct_weighted',
#  'smoothed_pct_contact_covid_positive_weighted',
#  'smoothed_pct_diabetes_weighted',
#  'smoothed_pct_cancer_weighted',
#  'smoothed_pct_heart_disease_weighted',
#  'smoothed_pct_high_blood_pressure_weighted',
#  'smoothed_pct_asthma_weighted',
#  'smoothed_pct_chronic_lung_disease_weighted',
#  'smoothed_pct_kidney_disease_weighted',
#  'smoothed_pct_autoimmune_disorder_weighted',
#  'smoothed_pct_no_above_medical_conditions_weighted',
#  'smoothed_pct_multiple_medical_conditions_weighted']
```

In [349]:
```
df1_tx_filtered= df1_tx.copy()

# df1_tx_filtered= df1_tx[columns_required]
df = df1_tx_filtered
df.head()
```

Out[349]:

date	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	smoothe
2020-11-01	tx	female	18-34	1367.0	0.8105	0.8114	
2020-10-31	tx	female	18-34	1418.0	0.8521	0.8529	
2020-10-30	tx	female	18-34	1457.0	1.0361	1.0369	
2020-10-29	tx	female	18-34	1586.0	1.2684	1.3324	
2020-10-28	tx	female	18-34	1706.0	1.2390	1.2985	

5 rows × 102 columns

In [ ]:
```
```

Exploratory analysis

In [350]:
```
df.groupby('gender').count()['summed_n']
```

Out[350]:
```
gender
female     816
male       816
overall    816
Name: summed_n, dtype: int64
```

In [351]:
```
fig = plt.figure()
gender =['female','male']


# df.groupby('gender').count()['date']
```

```
<Figure size 432x288 with 0 Axes>
```

Let's see what we have for particular date

In [352]:
```
df[df.index=='11/1/2020']
```

Out[352]:

date	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	smoothe
2020-11-01	tx	female	18-34	1367.0	0.8105	0.8114	
2020-11-01	tx	female	35-54	2536.0	1.1503	1.1502	
2020-11-01	tx	female	55+	3152.0	0.6378	0.5741	
2020-11-01	tx	female	overall	7055.0	0.8560	0.8275	
2020-11-01	tx	male	18-34	546.0	0.5566	0.7420	
2020-11-01	tx	male	35-54	1479.0	0.8822	0.9498	
2020-11-01	tx	male	55+	1796.0	0.6159	0.5600	
2020-11-01	tx	male	overall	3821.0	0.7114	0.7377	
2020-11-01	tx	overall	18-34	1913.0	0.7369	0.7905	
2020-11-01	tx	overall	35-54	4015.0	1.0517	1.0767	
2020-11-01	tx	overall	55+	4948.0	0.6300	0.5691	
2020-11-01	tx	overall	overall	10876.0	0.8052	0.7959	

12 rows × 102 columns

Here, we are going to see male and female in geneder bucket and do some analysis over overall age bucket

In [353]:
```
df_overall_gender =df[(df['age_bucket']=='overall') & (df['gender']!='overall'
)]
df_overall_gender.head(2)
```

Out[353]:

date	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	smoothe
2020-11-01	tx	female	overall	7055.0	0.8560	0.8275	
2020-10-31	tx	female	overall	7078.0	0.8815	0.8815	

2 rows × 102 columns

In [354]:
```
bins= [-1,0.000000001,1100]
labels = ['Negative','Postive']

df_overall_gender['covid_status'] = pd.cut(df_overall_gender['smoothed_pct_tes
ted_and_positive_weighted'], bins=bins, labels=labels, right=False)
```

In [355]:
```
df_overall_gender[['smoothed_pct_tested_and_positive_weighted','covid_status'
]].head(100)
```

Out[355]:

date	smoothed_pct_tested_and_positive_weighted	covid_status
2020-11-01	0.0000	Negative
2020-10-31	0.0000	Negative
2020-10-30	0.0000	Negative
2020-10-29	0.0000	Negative
2020-10-28	0.0000	Negative
...	...	...
2020-07-29	2.6698	Postive
2020-07-28	2.6368	Postive
2020-07-27	2.6605	Postive
2020-07-26	2.6207	Postive
2020-07-25	2.6329	Postive

100 rows × 2 columns

In [356]:
```
df_overall_gender.groupby('covid_status').count()
```

Out[356]:

covid_status	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	s
Negative	97	97	97	97	97	97	
Postive	311	311	311	311	311	311	

2 rows × 102 columns

In [357]:
```
df_overall_gender[df_overall_gender['smoothed_pct_tested_and_positive_weighted']>0]
```

Out[357]:

date	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	smoothe
2020-09-14	tx	female	overall	12333.0	0.7106	0.7024	
2020-09-13	tx	female	overall	12178.0	0.6703	0.6620	
2020-09-12	tx	female	overall	11837.0	0.6303	0.6388	
2020-09-11	tx	female	overall	11346.0	0.6579	0.6668	
2020-09-10	tx	female	overall	11033.0	0.6771	0.7228	
...	...	...	...	...	...	...	
2020-04-16	tx	male	overall	5858.0	0.7291	0.6941	
2020-04-15	tx	male	overall	5951.0	0.8367	0.7853	
2020-04-14	tx	male	overall	6318.0	0.7730	0.7569	
2020-04-13	tx	male	overall	6357.0	0.8323	0.8163	
2020-04-12	tx	male	overall	6376.0	0.8462	0.8143	

311 rows × 103 columns

Let's look at patients with covid-like illness and influenza-like illness over the time in 2020

Spring: March-May (03 - 05)

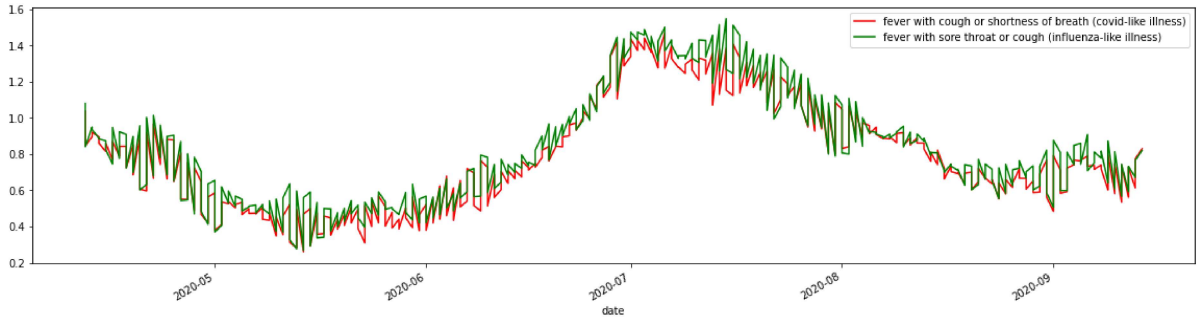
Summer: June-August (06 - 07)

Autumn: September-November(08 - 11)

Winter: December-Feb(12 - 2)

In [358]:
```
df_overall_gender_positive_test= df_overall_gender[df_overall_gender['covid_status']=='Postive']
```

In [359]:
```
fig = plt.figure(figsize = (20, 5))

df_overall_gender_positive_test['smoothed_pct_cli_weighted'].plot(label='fever with cough or shortness of breath (covid-like illness)',color='red')
df_overall_gender_positive_test['smoothed_pct_ili_weighted'].plot(label='fever with sore throat or cough (influenza-like illness)',color='green')

plt.legend()
plt.show()


# df_overall_gender[df_overall_gender['covid_status']=='Postive']['smoothed_pct_cli_weighted'].plot(label='fever with cough or shortness of breath (covid-like illness)')
```



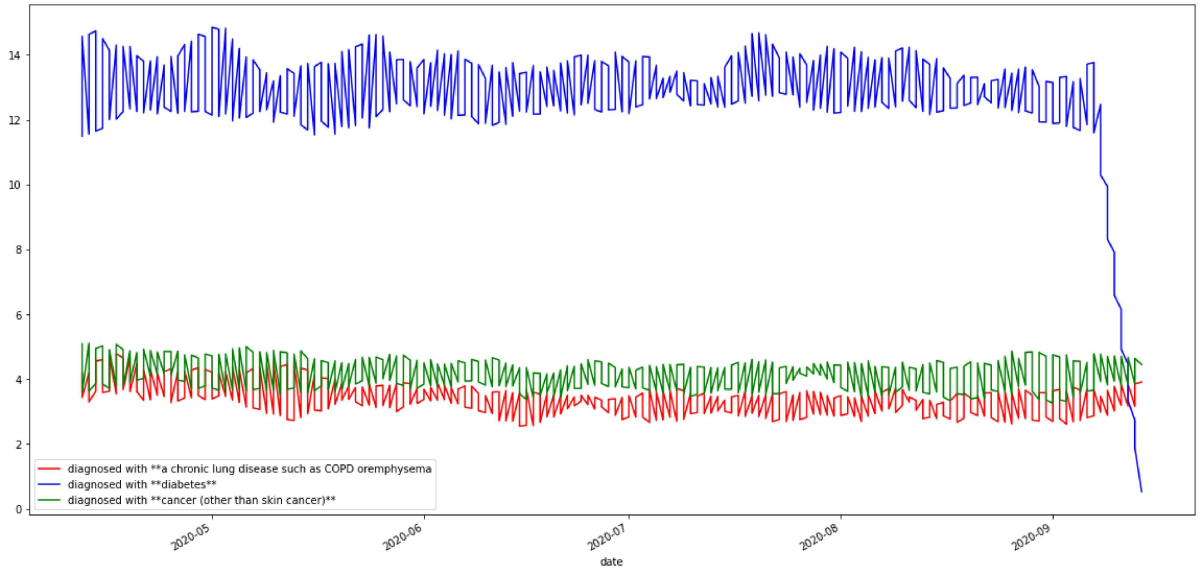
The prevalence of diabetes has gradually increased over the past 10 years, both in Texas and nationally (1). In 2017, an estimated 2,323,220 people in Texas had diabetes, which represented 11.4% of the adult population (2). Additionally, approximately 23.8% of people who had diabetes were not aware of it (3).

Source:
https://www.cdc.gov/pcd/issues/2019/19_0175.htm#:~:text=The%20prevalence%20of%20diabetes%20has,aware%
(https://www.cdc.gov/pcd/issues/2019/19_0175.htm#:~:text=The%20prevalence%20of%20diabetes%20has,aware

About 5.4% (age-adjusted = 5.5%) of Texas residents surveyed in 2011 reported having been told by a health care professional that they have COPD. The map below depicts quartiles of the national prevalence of COPD by state for comparison

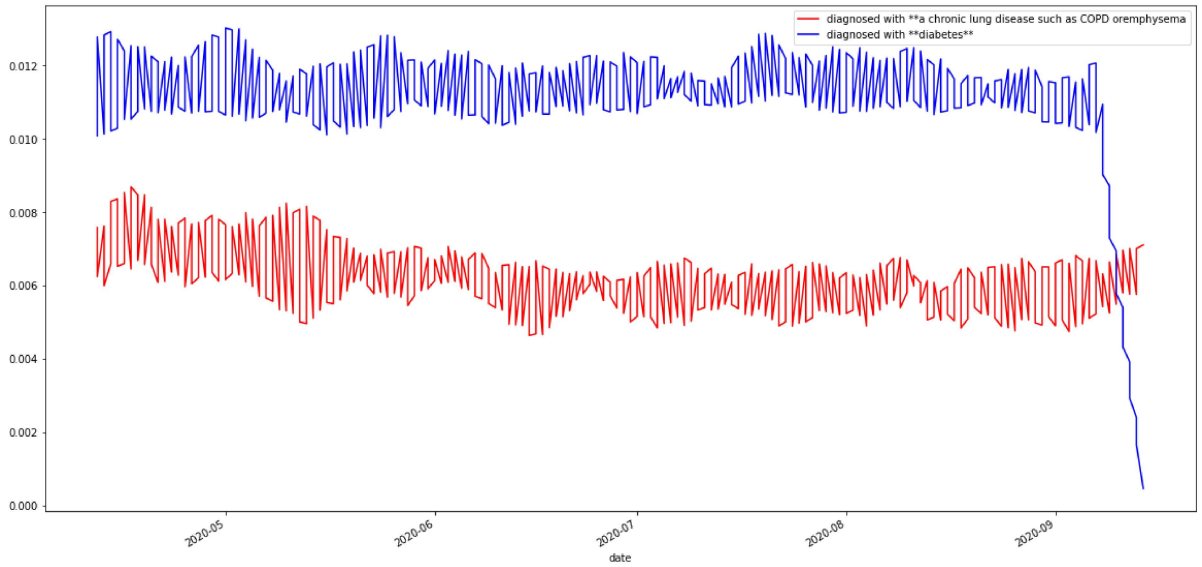
https://www.cdc.gov/copd/maps/docs/pdf/TX_COPDFactSheet.pdf
(https://www.cdc.gov/copd/maps/docs/pdf/TX_COPDFactSheet.pdf)

In [360]:
```
fig = plt.figure(figsize = (20, 10))
df_overall_gender_positive_test['smoothed_pct_chronic_lung_disease_weighted'].
plot(label=

'diagnosed with **a chronic lung disease such as COPD oremphysema',color='red'
)
df_overall_gender_positive_test['smoothed_pct_diabetes_weighted'].plot(label=

'diagnosed with **diabetes**',color='blue')
df_overall_gender_positive_test['smoothed_pct_cancer_weighted'].plot(label=

'diagnosed with **cancer (other than skin cancer)**',color='green')



plt.legend()
plt.show()
```



In [361]:
```
df_overall_gender_positive_test['adjusted_smoothed_pct_chronic_lung_disease_we
ighted']=df_overall_gender_positive_test['smoothed_pct_chronic_lung_disease_we
ighted']/(5.5*100)

df_overall_gender_positive_test['adjusted_smoothed_pct_diabetes_weighted']=df_
overall_gender_positive_test['smoothed_pct_diabetes_weighted']/(11.4*100)
```

In [362]:
```
fig = plt.figure(figsize = (20, 10))

df_overall_gender_positive_test['adjusted_smoothed_pct_chronic_lung_disease_we
ighted'].plot(label=

'diagnosed with **a chronic lung disease such as COPD oremphysema',color='red'
)
df_overall_gender_positive_test['adjusted_smoothed_pct_diabetes_weighted'].plo
t(label=

'diagnosed with **diabetes**',color='blue')
# df_overall_gender_positive_test['smoothed_pct_cancer_weighted'].plot(label=
#
'diagnosed with **cancer (other than skin cancer)**',color='green')



plt.legend()
plt.show()
```



32.5 percent has high blood pressure https://www.americashealthrankings.org/explore/annual/measure/Hypertension/state/TX (https://www.americashealthrankings.org/explore/annual/measure/Hypertension/state/TX)

Heart disease and stroke are the number one and number three causes of death in Texas.1 Together, they account for nearly three of every ten deaths in the state https://www.dshs.state.tx.us/heart/pdf/Texas-Public-Health-Strategies_CVDS-2019-2023-Final.pdf (https://www.dshs.state.tx.us/heart/pdf/Texas-Public-Health-Strategies_CVDS-2019-2023-Final.pdf)

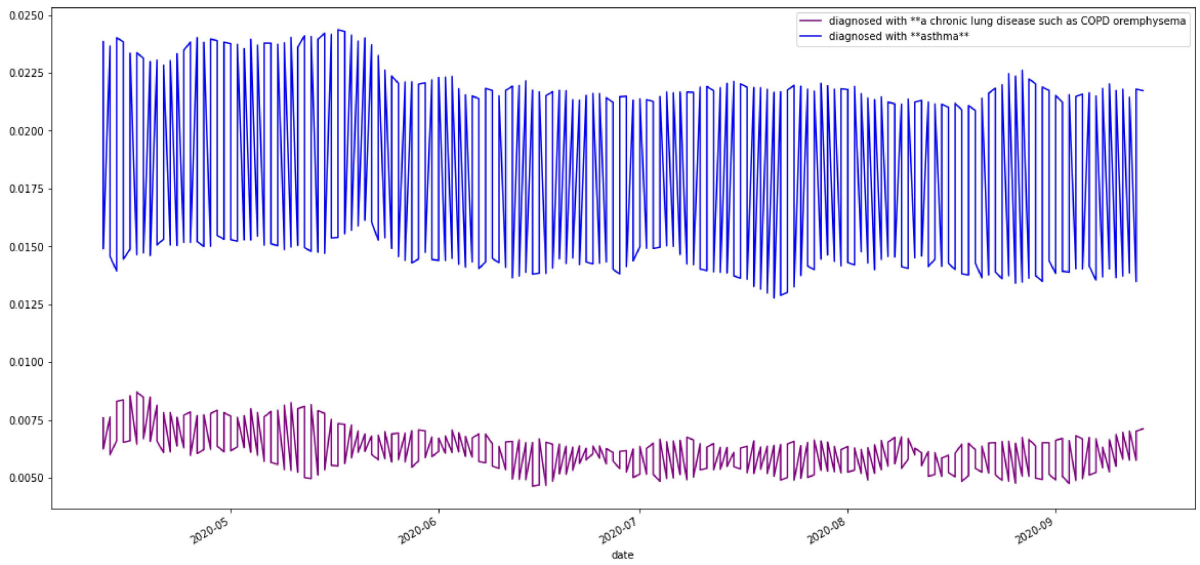
7.1 percent has asthma https://www.cdc.gov/asthma/most_recent_data_states.htm (https://www.cdc.gov/asthma/most_recent_data_states.htm)

7 percent has autoimmune disorder https://www.gene.com/stories/autoimmune-disease-101 (https://www.gene.com/stories/autoimmune-disease-101)

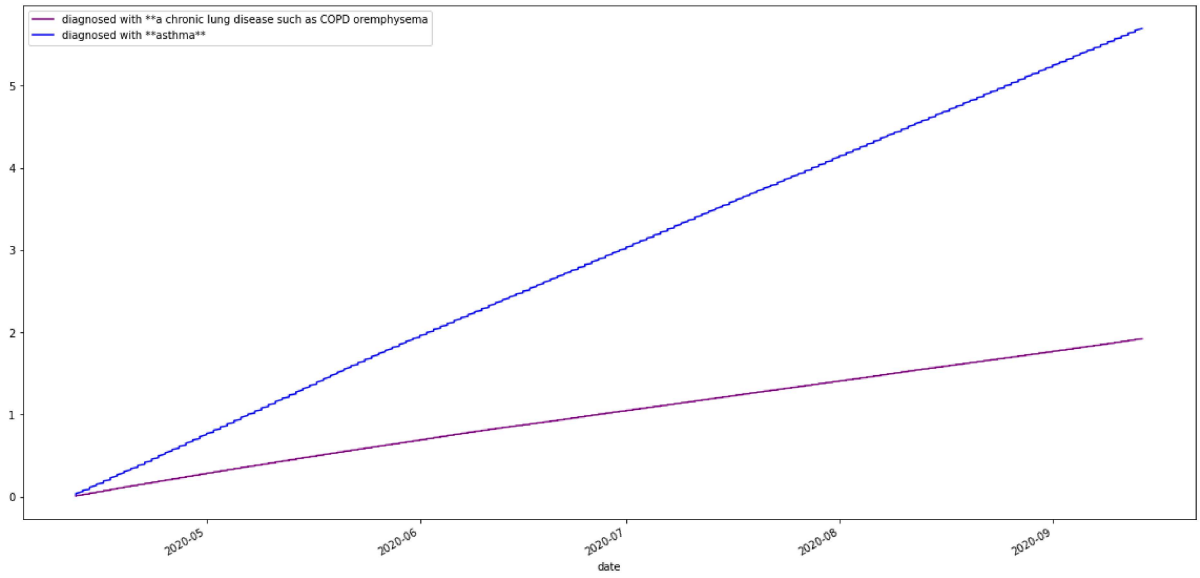
In [363]:
```
df_overall_gender_positive_test['adjusted_smoothed_pct_high_blood_pressure_weighted']=df_overall_gender_positive_test['smoothed_pct_high_blood_pressure_weighted']/(32.5*100)

df_overall_gender_positive_test['adjusted_smoothed_pct_heart_disease_weighted']=df_overall_gender_positive_test['smoothed_pct_heart_disease_weighted']/(30*100)

df_overall_gender_positive_test['adjusted_smoothed_pct_asthma_weighted']=df_overall_gender_positive_test['smoothed_pct_asthma_weighted']/(7.1*100)

df_overall_gender_positive_test['adjusted_smoothed_pct_autoimmune_disorder_weighted']=df_overall_gender_positive_test['smoothed_pct_autoimmune_disorder_weighted']/(7*100)
```

In [364]:
```
fig = plt.figure(figsize = (20, 10))
# df_overall_gender_positive_test['smoothed_pct_autoimmune_disorder_weighte
d'].plot(label=
#
'diagnosed with **an autoimmune disorder such as rheumatoid arthritis or Crohn
s disease**',color='red')
# df_overall_gender_positive_test['smoothed_pct_asthma_weighted'].plot(label=
#
'diagnosed with **asthma**',color='blue')
# df_overall_gender_positive_test['smoothed_pct_heart_disease_weighted'].plot
(label=
#
'diagnosed with heart disease',color='green')

# df_overall_gender_positive_test['smoothed_pct_high_blood_pressure_weighte
d'].plot(label=
#
'diagnosed with kidney disease',color='black')

# plt.legend()
# plt.show()

fig = plt.figure(figsize = (20, 10))

df_overall_gender_positive_test['adjusted_smoothed_pct_chronic_lung_disease_we
ighted'].plot(label=

'diagnosed with **a chronic lung disease such as COPD oremphysema',color='purp
le')

df_overall_gender_positive_test['adjusted_smoothed_pct_asthma_weighted'].plot(
label=

'diagnosed with **asthma**',color='blue')

plt.legend()
plt.show()
```

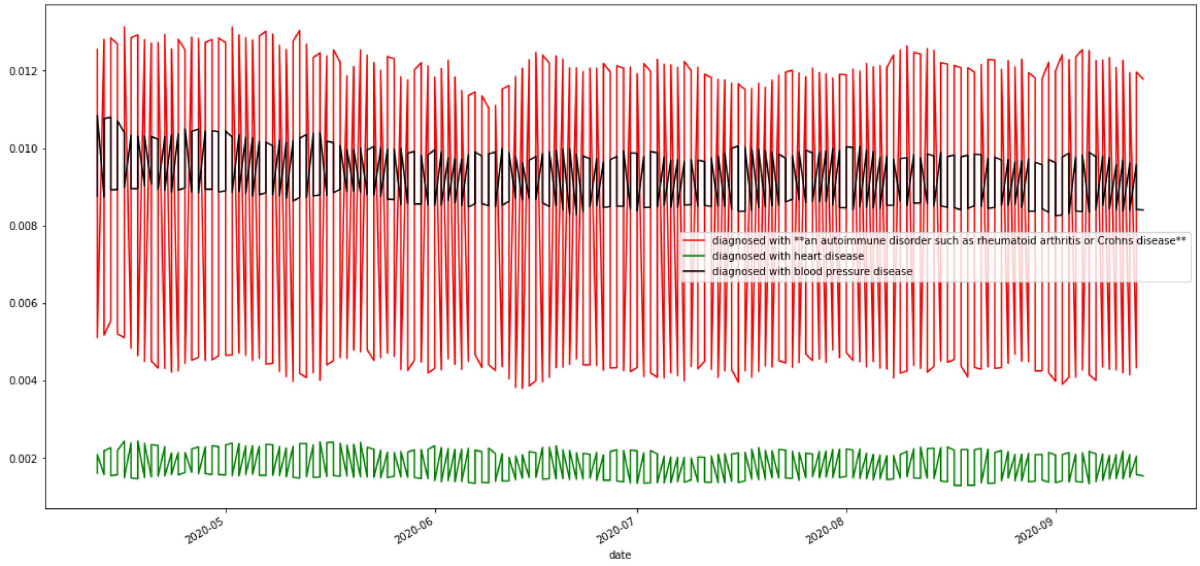
```
<Figure size 1440x720 with 0 Axes>
```



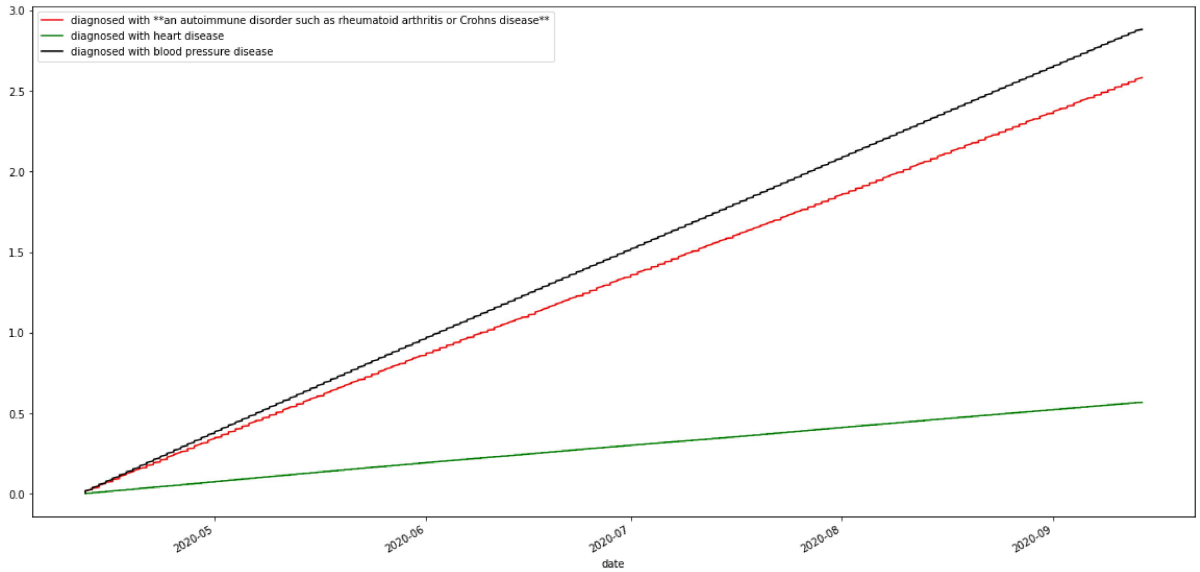
In [365]:
```
fig = plt.figure(figsize = (20, 10))

df_overall_gender_positive_test['adjusted_smoothed_pct_chronic_lung_disease_weighted'].sort_index().cumsum().plot(label=

'diagnosed with **a chronic lung disease such as COPD oremphysema',color='purple')

df_overall_gender_positive_test['adjusted_smoothed_pct_asthma_weighted'].sort_index().cumsum().plot(label=

'diagnosed with **asthma**',color='blue')

plt.legend()
plt.show()
```



In [366]:
```
fig = plt.figure(figsize = (20, 10))

df_overall_gender_positive_test['adjusted_smoothed_pct_autoimmune_disorder_weighted'].plot(label=

'diagnosed with **an autoimmune disorder such as rheumatoid arthritis or Crohns disease**',color='red')

df_overall_gender_positive_test['adjusted_smoothed_pct_heart_disease_weighted'].plot(label=

'diagnosed with heart disease',color='green')

df_overall_gender_positive_test['adjusted_smoothed_pct_high_blood_pressure_weighted'].plot(label=

'diagnosed with blood pressure disease',color='black')
plt.legend()
plt.show()
```

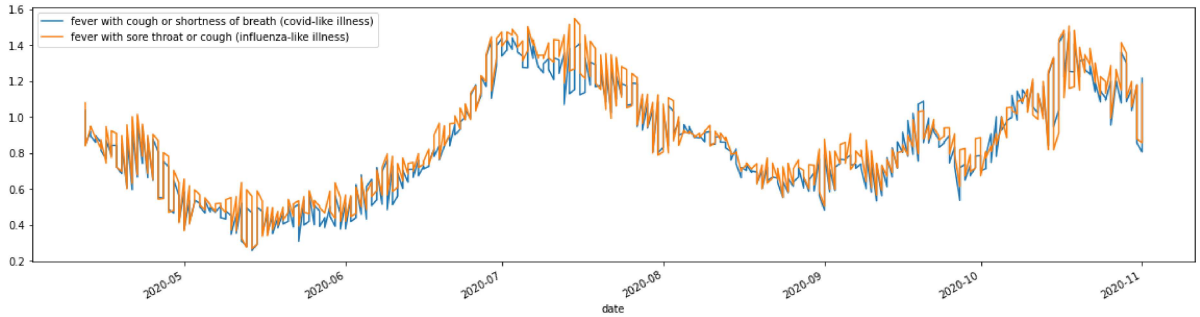


In [367]:
```
fig = plt.figure(figsize = (20, 10))

df_overall_gender_positive_test['adjusted_smoothed_pct_autoimmune_disorder_weighted'].sort_index().cumsum().plot(label=

'diagnosed with **an autoimmune disorder such as rheumatoid arthritis or Crohns disease**',color='red')

df_overall_gender_positive_test['adjusted_smoothed_pct_heart_disease_weighted'].sort_index().cumsum().plot(label=

'diagnosed with heart disease',color='green')

df_overall_gender_positive_test['adjusted_smoothed_pct_high_blood_pressure_weighted'].sort_index().cumsum().plot(label=

'diagnosed with blood pressure disease',color='black')
plt.legend()
plt.show()
```



In [ ]:
```
```

In [368]:
```
fig = plt.figure(figsize = (20, 5))

df_overall_gender['smoothed_pct_cli_weighted'].plot(label='fever with cough or shortness of breath (covid-like illness)')
df_overall_gender['smoothed_pct_ili_weighted'].plot(label='fever with sore throat or cough (influenza-like illness)')

plt.legend()
plt.show()
```



In [ ]:
```
```

Estimated percentage of people with COVID-like illness, which is defined as exhibiting a fever along with cough or shortness of breath or difficulty breathing.

In [376]:
```
# # To generate a plot showing trend of 'smoothed_pct_cli_weighted' over the given time series.
# import matplotlib.pyplot as plt

# fig = plt.figure(figsize = (20, 5))

# plt.plot('date','smoothed_pct_cli_weighted', data= df.sort_values('date'))
# plt.show()
```

In [378]:
```
df.dtypes
```

Out[378]:
```
state_code                                          object
gender                                              object
age_bucket                                          object
summed_n                                           float64
smoothed_pct_cli                                   float64
                                                    ...
smoothed_pct_chronic_lung_disease_weighted         float64
smoothed_pct_kidney_disease_weighted               float64
smoothed_pct_autoimmune_disorder_weighted          float64
smoothed_pct_no_above_medical_conditions_weighted  float64
smoothed_pct_multiple_medical_conditions_weighted  float64
Length: 102, dtype: object
```

In [380]:
```
# To convert data type into datetime
df['date'] = pd.to_datetime(df.index)

# plt.plot('date','smoothed_pct_cli_weighted', data= df.sort_values('date'))
```

In [381]:
```
df.dtypes
```

Out[381]:
```
state_code                                              object
gender                                                  object
age_bucket                                              object
summed_n                                               float64
smoothed_pct_cli                                       float64
                                                          ...
smoothed_pct_kidney_disease_weighted                   float64
smoothed_pct_autoimmune_disorder_weighted              float64
smoothed_pct_no_above_medical_conditions_weighted      float64
smoothed_pct_multiple_medical_conditions_weighted      float64
date                                            datetime64[ns]
Length: 103, dtype: object
```

In [382]:
```
df.head()
```

Out[382]:

date	state_code	gender	age_bucket	summed_n	smoothed_pct_cli	smoothed_pct_ili	smoothe
2020-11-01	tx	female	18-34	1367.0	0.8105	0.8114	
2020-10-31	tx	female	18-34	1418.0	0.8521	0.8529	
2020-10-30	tx	female	18-34	1457.0	1.0361	1.0369	
2020-10-29	tx	female	18-34	1586.0	1.2684	1.3324	
2020-10-28	tx	female	18-34	1706.0	1.2390	1.2985	

5 rows × 103 columns

In [392]:
```
# Since the gender has 'overall" rows, we want to eliminate those rows.
df.index.names = ['Date2']
df = df[df['gender']!='overall']
df.shape
```

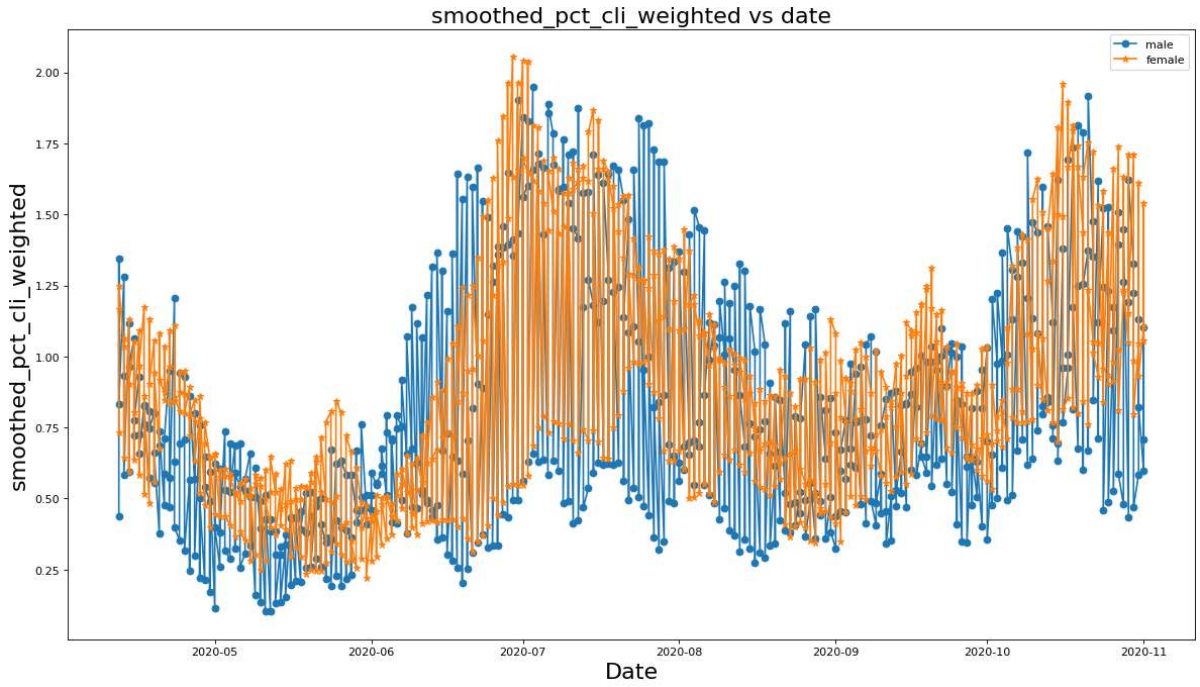
Out[392]:
```
(1224, 103)
```

In [393]:
```
df['gender'].value_counts()
```

Out[393]:
```
male      612
female    612
Name: gender, dtype: int64
```

In [394]:
```
df_male = df[df['gender']=='male']
df_female = df[df['gender']=='female']
```

In [395]:
```
plt.figure(num=None, figsize=(18, 10), dpi=80, facecolor='w', edgecolor='k')
plt.plot('date','smoothed_pct_cli_weighted', marker='o',data= df_male.sort_val
ues('date'), label='male')
plt.plot('date','smoothed_pct_cli_weighted', marker='*',data= df_female.sort_v
alues('date'),label='female')
plt.title('smoothed_pct_cli_weighted vs date', fontdict={'fontsize': 20})
plt.xlabel('Date', fontdict={'fontsize': 20})
plt.ylabel('smoothed_pct_cli_weighted', fontdict={'fontsize': 20})
plt.legend()
plt.show()
```

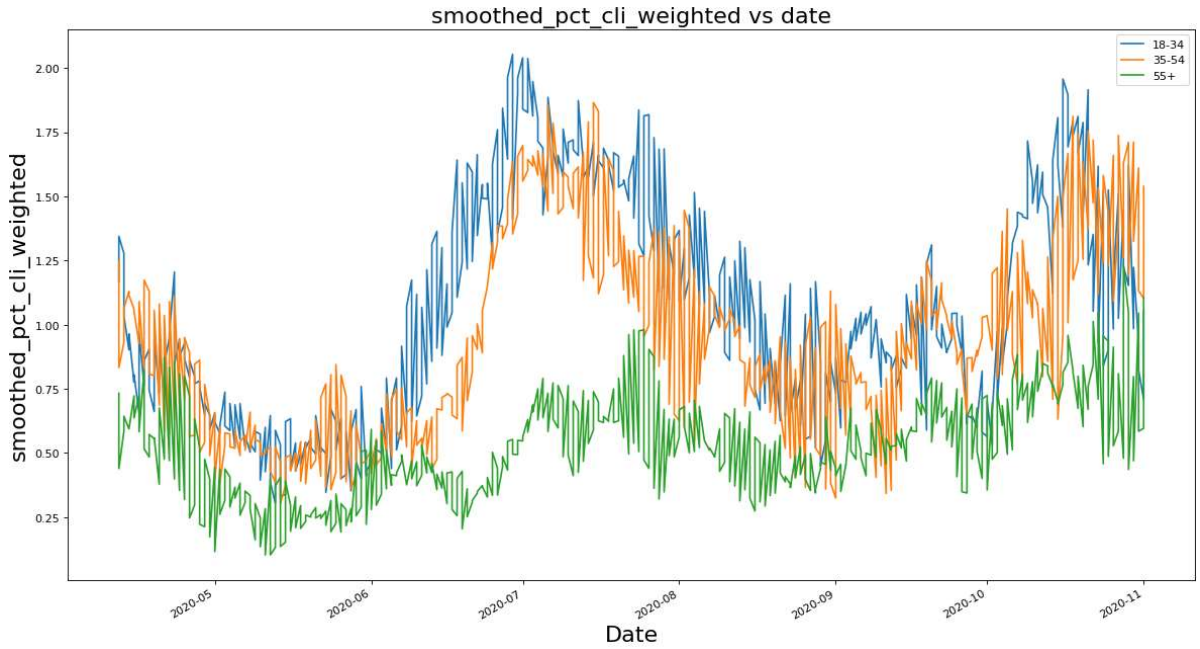


In [396]:
```
# It seems that comparatively female population with cli are less overall, pro
bably females are more cautious and take more protective measures than males?
```

In [397]:
```
# Another way to generate the plot using x-axis as date
df =df.set_index('date')
```

In [398]:
```
plt.figure(num=None, figsize=(18, 10), dpi=80, facecolor='w', edgecolor='k')
df.groupby('age_bucket')['smoothed_pct_cli_weighted'].plot(legend=True)
plt.title('smoothed_pct_cli_weighted vs date', fontdict={'fontsize': 20})
plt.xlabel('Date', fontdict={'fontsize': 20})
plt.ylabel('smoothed_pct_cli_weighted', fontdict={'fontsize': 20})
# plt.savefig('test.png', dpi=300)
```

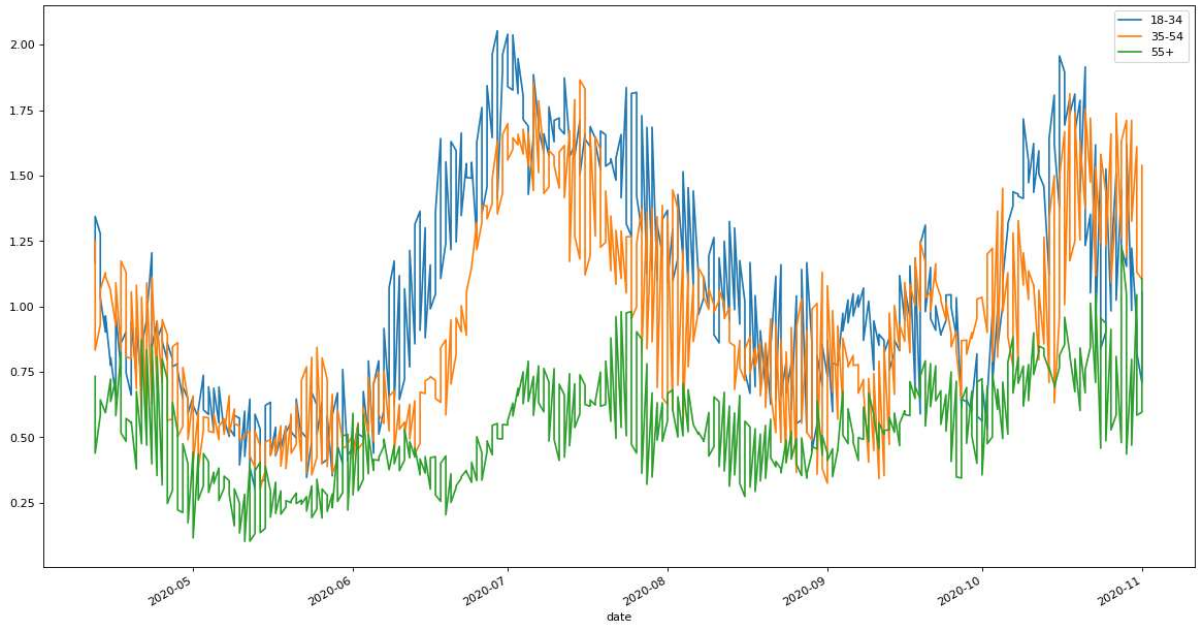
Out[398]:
```
Text(0, 0.5, 'smoothed_pct_cli_weighted')
```



In [399]:
```
# let's exclude 'overall' from the age-bucket
df = df[df['age_bucket']!='overall']
```

In [400]:
```
plt.figure(num=None, figsize=(18, 10), dpi=80, facecolor='w', edgecolor='k')
df.groupby('age_bucket')['smoothed_pct_cli_weighted'].plot(legend=True)
```

Out[400]:
```
age_bucket
18-34    AxesSubplot(0.125,0.2;0.775x0.68)
35-54    AxesSubplot(0.125,0.2;0.775x0.68)
55+      AxesSubplot(0.125,0.2;0.775x0.68)
Name: smoothed_pct_cli_weighted, dtype: object
```



In [401]:
```
# It appears that all age_bucket had similar incidence of cli in the begining
 of May, but over the months age_bucket 18-34 have higher number of cli, possi
bly because these are young people and did not take sufficient precautions.We
 can see that there was a spkike during the month of July probably because of
 ....
```