BIL 366 Data Mining: Homework-1

Soru1:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

url = 'https://drive.google.com/file/d/18gyHbx6rfogq3yQ-
GR9COjcGgyYlCnBZ/view?usp=sharing'
url2 = 'https://drive.google.com/uc?id=' + url.split('/')[-2]
df = pd.read_csv(url2, usecols=['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'])
df.info()

df.describe().iloc[3:]
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167657 entries, 0 to 167656
Data columns (total 7 columns):
 #   Column                                              Non-Null
Count   Dtype
---  ------
--------------   -----
 0   date                                                167657 non-
null  object
 1   retail_and_recreation_percent_change_from_baseline  101865 non-
null  float64
 2   grocery_and_pharmacy_percent_change_from_baseline   106104 non-
null  float64
 3   parks_percent_change_from_baseline                  95186 non-
null   float64
 4   transit_stations_percent_change_from_baseline       87723 non-
null   float64
 5   workplaces_percent_change_from_baseline             158870 non-
null  float64
 6   residential_percent_change_from_baseline            98651 non-
null   float64
dtypes: float64(6), object(1)
memory usage: 9.0+ MB

     retail_and_recreation_percent_change_from_baseline  \
min                                                -100.0
25%                                                 -44.0
```

```
50%                                              -24.0
75%                                               -8.0
max                                              333.0

     grocery_and_pharmacy_percent_change_from_baseline  \
min                                             -100.0
25%                                               -9.0
50%                                                5.0
75%                                               18.0
max                                              321.0

     parks_percent_change_from_baseline  \
min                              -100.0
25%                               -26.0
50%                                 2.0
75%                                30.0
max                               694.0

     transit_stations_percent_change_from_baseline  \
min                                         -100.0
25%                                          -48.0
50%                                          -25.0
75%                                           -5.0
max                                          318.0

     workplaces_percent_change_from_baseline  \
min                                    -94.0
25%                                    -30.0
50%                                    -17.0
75%                                     -6.0
max                                    136.0

     residential_percent_change_from_baseline
min                                     -28.0
25%                                       1.0
50%                                       5.0
75%                                      12.0
max                                      50.0
```

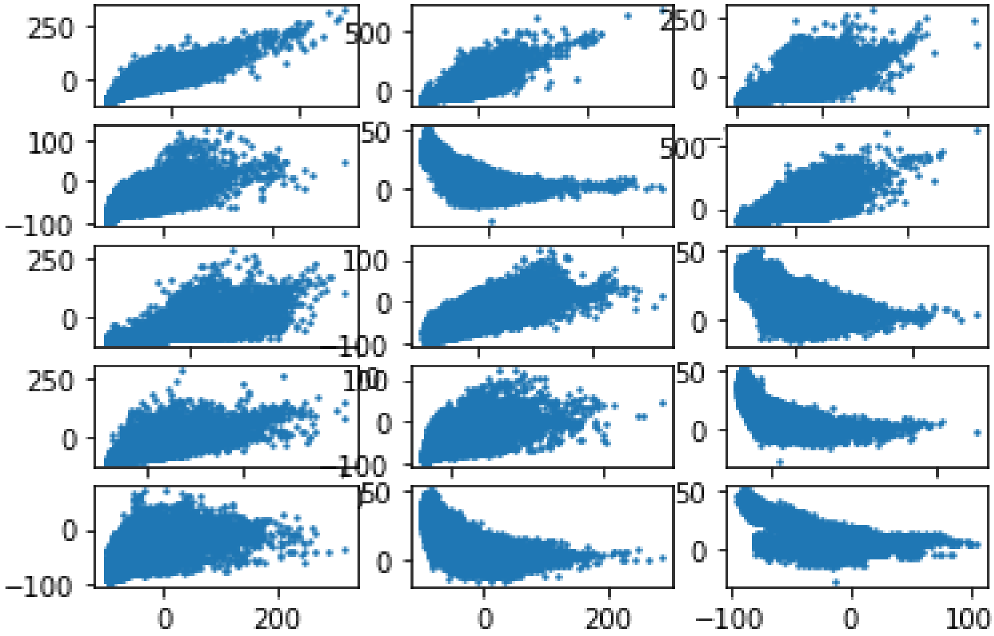
Soru2:

```
fig, axs = plt.subplots(5, 3)

data1 = df.retail_and_recreation_percent_change_from_baseline
data2 = df.grocery_and_pharmacy_percent_change_from_baseline
data3 = df.parks_percent_change_from_baseline
data4 = df.transit_stations_percent_change_from_baseline
data5 = df.workplaces_percent_change_from_baseline
data6 = df.residential_percent_change_from_baseline

axs[0, 0].scatter(data1, data2, s=2)
```

```
axs[0, 1].scatter(data1, data3, s=2)
axs[0, 2].scatter(data1, data4, s=2)
axs[1, 0].scatter(data1, data5, s=2)
axs[1, 1].scatter(data1, data6, s=2)
axs[1, 2].scatter(data2, data3, s=2)
axs[2, 0].scatter(data2, data4, s=2)
axs[2, 1].scatter(data2, data5, s=2)
axs[2, 2].scatter(data2, data6, s=2)
axs[3, 0].scatter(data3, data4, s=2)
axs[3, 1].scatter(data3, data5, s=2)
axs[3, 2].scatter(data3, data6, s=2)
axs[4, 0].scatter(data4, data5, s=2)
axs[4, 1].scatter(data4, data6, s=2)
axs[4, 2].scatter(data5, data6, s=2)

data = df.corr(method='pearson')
np.sign(data)
```

```
retail_and_recreation_percent_change_from_baseline  \
retail_and_recreation_percent_change_from_baseline
1.0
grocery_and_pharmacy_percent_change_from_baseline
1.0
parks_percent_change_from_baseline
1.0
transit_stations_percent_change_from_baseline
1.0
workplaces_percent_change_from_baseline
1.0
residential_percent_change_from_baseline
-1.0


grocery_and_pharmacy_percent_change_from_baseline  \
retail_and_recreation_percent_change_from_baseline
1.0
grocery_and_pharmacy_percent_change_from_baseline
1.0
parks_percent_change_from_baseline
1.0
transit_stations_percent_change_from_baseline
1.0
workplaces_percent_change_from_baseline
1.0
residential_percent_change_from_baseline
-1.0


parks_percent_change_from_baseline  \
retail_and_recreation_percent_change_from_baseline
1.0
grocery_and_pharmacy_percent_change_from_baseline
1.0
parks_percent_change_from_baseline
1.0
transit_stations_percent_change_from_baseline
1.0
workplaces_percent_change_from_baseline
1.0
residential_percent_change_from_baseline
-1.0


transit_stations_percent_change_from_baseline  \
retail_and_recreation_percent_change_from_baseline
1.0
grocery_and_pharmacy_percent_change_from_baseline
1.0
parks_percent_change_from_baseline
1.0
transit_stations_percent_change_from_baseline
1.0
workplaces_percent_change_from_baseline
1.0
residential_percent_change_from_baseline
-1.0


workplaces_percent_change_from_baseline  \
retail_and_recreation_percent_change_from_baseline
1.0
grocery_and_pharmacy_percent_change_from_baseline
1.0
parks_percent_change_from_baseline
1.0
transit_stations_percent_change_from_baseline
1.0
workplaces_percent_change_from_baseline
1.0
residential_percent_change_from_baseline
-1.0


residential_percent_change_from_baseline
retail_and_recreation_percent_change_from_baseline
-1.0
grocery_and_pharmacy_percent_change_from_baseline
-1.0
parks_percent_change_from_baseline
```

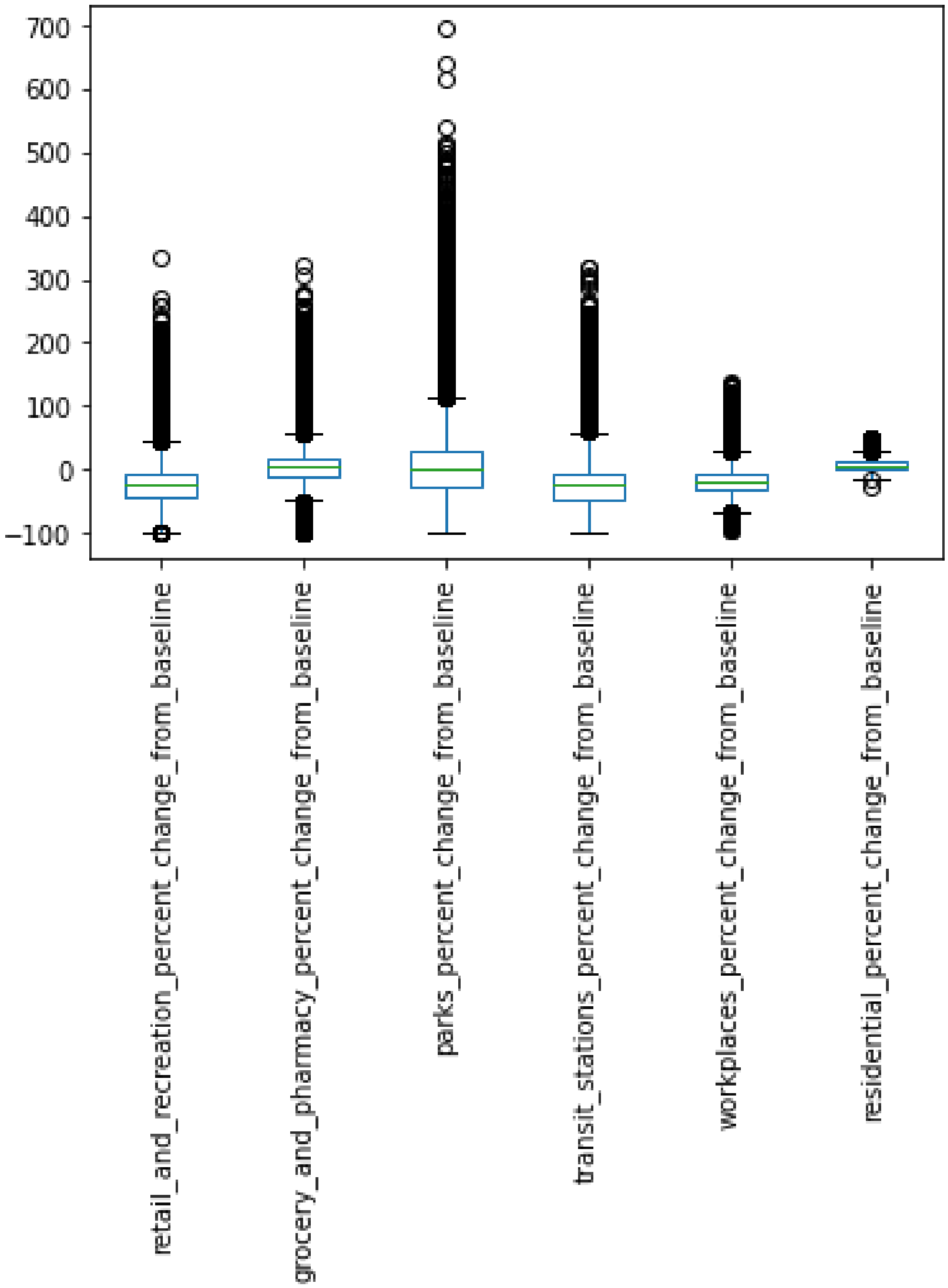
```
-1.0
transit_stations_percent_change_from_baseline
-1.0
workplaces_percent_change_from_baseline
-1.0
residential_percent_change_from_baseline
1.0
```



Soru3:

```
df.boxplot(column=['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'], rot=90, grid=False)
```

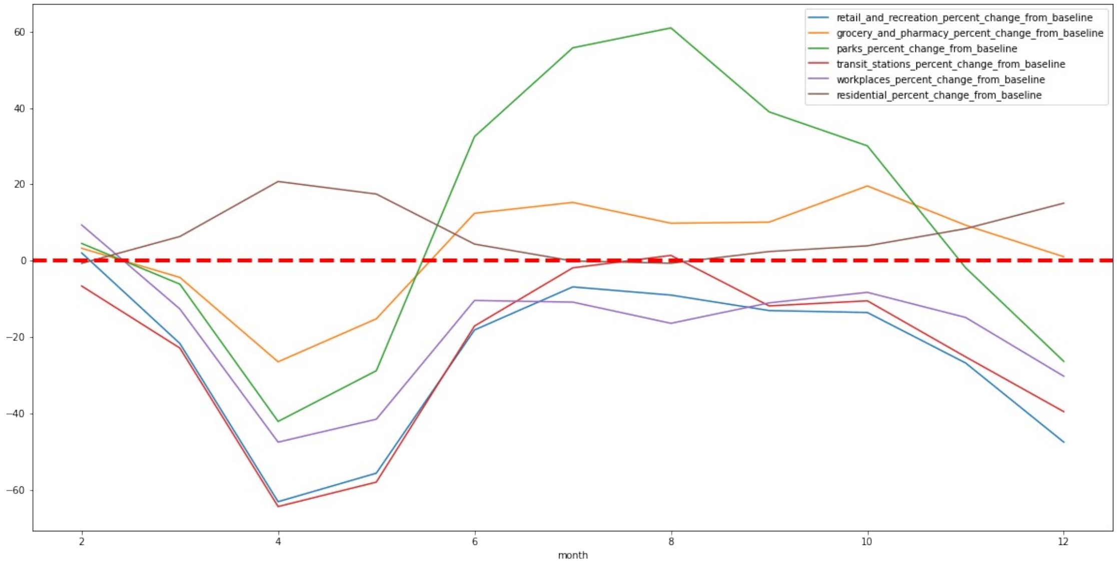
```
<AxesSubplot:>
```



Soru 4:

```
df['date'] = pd.to_datetime(df['date'])
df['month'] = pd.DatetimeIndex(df['date']).month

data2020 = df.groupby(df.date.dt.to_period("M")).mean()

data2020.plot(x='month' ,figsize=(20, 10)).axhline(linewidth=4,
color='r', linestyle='--')
```

```
<matplotlib.lines.Line2D at 0x117cde370>
```



Soru5:

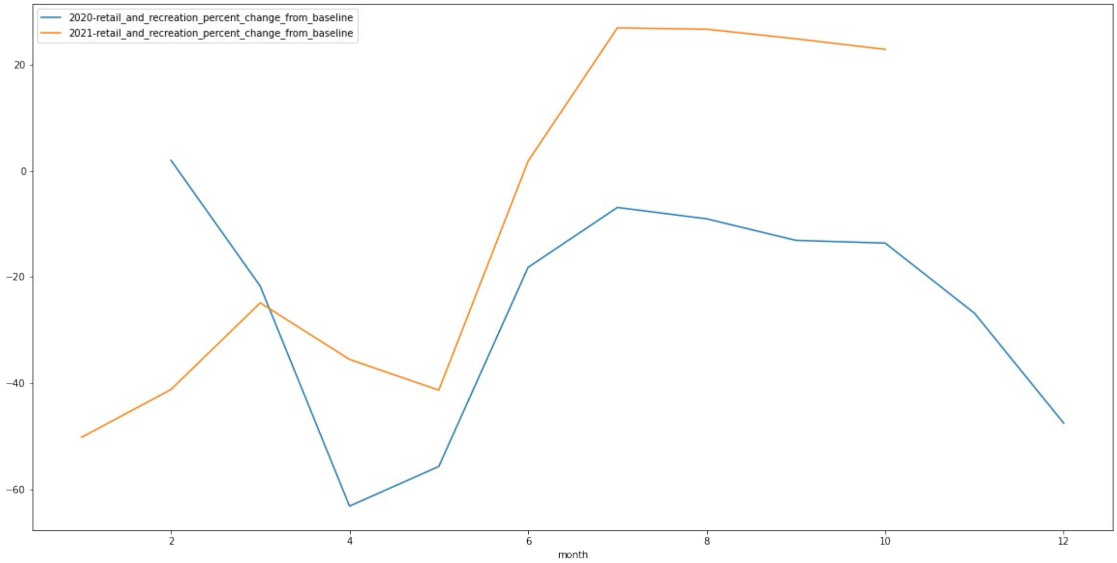
```
data2020 =
data2020.rename(columns={'retail_and_recreation_percent_change_from_ba
seline':'2020-retail_and_recreation_percent_change_from_baseline',
'grocery_and_pharmacy_percent_change_from_baseline':'2020-
grocery_and_pharmacy_percent_change_from_baseline',
'parks_percent_change_from_baseline':'2020-
parks_percent_change_from_baseline',
'transit_stations_percent_change_from_baseline':'2020-
transit_stations_percent_change_from_baseline',
'workplaces_percent_change_from_baseline':'2020-
workplaces_percent_change_from_baseline',
'residential_percent_change_from_baseline':'2020-
residential_percent_change_from_baseline'})

url3 = 'https://drive.google.com/file/d/1Eg8Lffm49bc-
bGFkv_4ddrQw8U8WE6P4/view?usp=sharing'
url4 = 'https://drive.google.com/uc?id=' + url3.split('/')[-2]
df2 = pd.read_csv(url4, usecols=['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'])
df2.info()


df2['date'] = pd.to_datetime(df2['date'])
df2['month'] = pd.DatetimeIndex(df2['date']).month
```

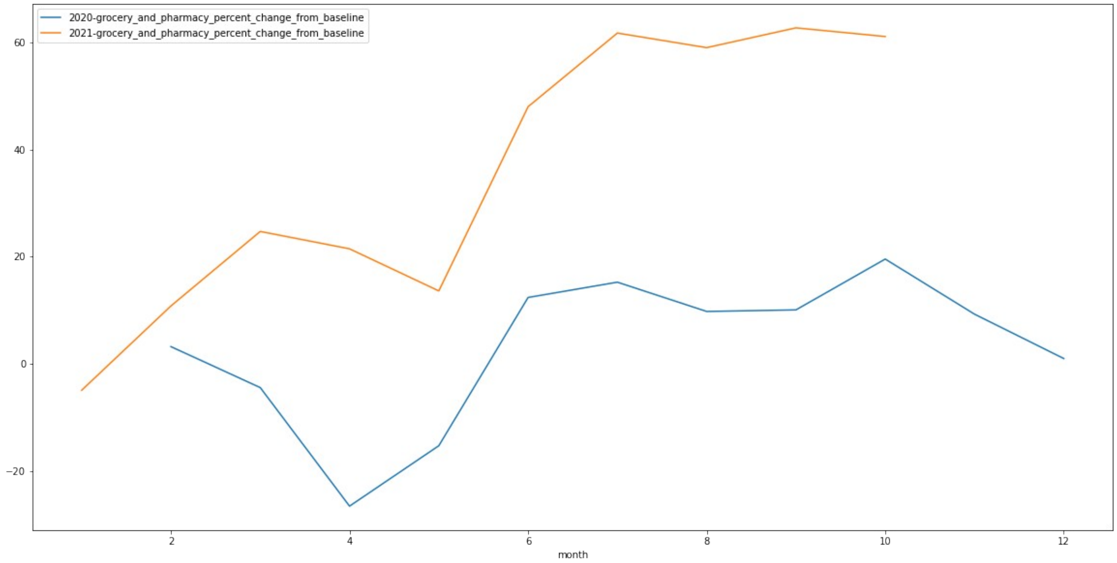
```
data2021 = df2.groupby(df2.date.dt.to_period('M')).mean()

data2021 =
data2021.rename(columns={'retail_and_recreation_percent_change_from_ba
seline':'2021-retail_and_recreation_percent_change_from_baseline',
'grocery_and_pharmacy_percent_change_from_baseline':'2021-
grocery_and_pharmacy_percent_change_from_baseline',
'parks_percent_change_from_baseline':'2021-
parks_percent_change_from_baseline',
'transit_stations_percent_change_from_baseline':'2021-
transit_stations_percent_change_from_baseline',
'workplaces_percent_change_from_baseline':'2021-
workplaces_percent_change_from_baseline',
'residential_percent_change_from_baseline':'2021-
residential_percent_change_from_baseline'})

calc = pd.concat([data2020, data2021])

calc.plot(x='month', y=['2020-
retail_and_recreation_percent_change_from_baseline','2021-
retail_and_recreation_percent_change_from_baseline'], figsize=(20,
10))
calc.plot(x='month', y=['2020-
grocery_and_pharmacy_percent_change_from_baseline','2021-
grocery_and_pharmacy_percent_change_from_baseline'], figsize=(20, 10))
calc.plot(x='month', y=['2020-
parks_percent_change_from_baseline','2021-
parks_percent_change_from_baseline'], figsize=(20, 10))
calc.plot(x='month', y=['2020-
transit_stations_percent_change_from_baseline','2021-
transit_stations_percent_change_from_baseline'], figsize=(20, 10))
calc.plot(x='month', y=['2020-
workplaces_percent_change_from_baseline','2021-
workplaces_percent_change_from_baseline'], figsize=(20, 10))
calc.plot(x='month', y=['2020-
residential_percent_change_from_baseline','2021-
residential_percent_change_from_baseline'], figsize=(20, 10))
```

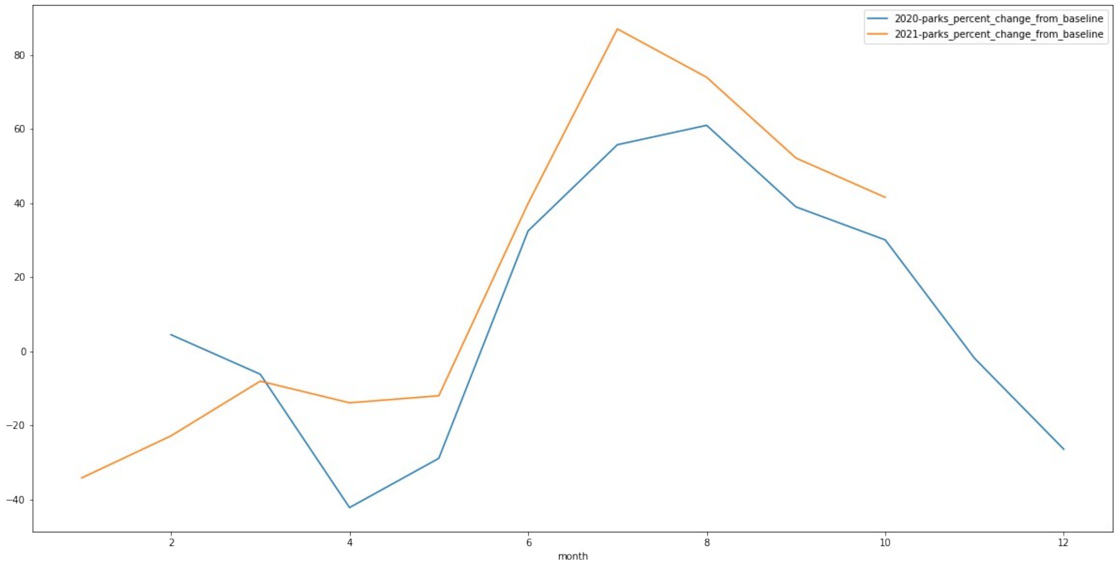
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158430 entries, 0 to 158429
Data columns (total 7 columns):
 #   Column                                              Non-Null
Count   Dtype
---  ------
--------------   -----
 0   date                                                158430 non-
null  object
 1   retail_and_recreation_percent_change_from_baseline  91170 non-
null   float64
 2   grocery_and_pharmacy_percent_change_from_baseline   92489 non-
```

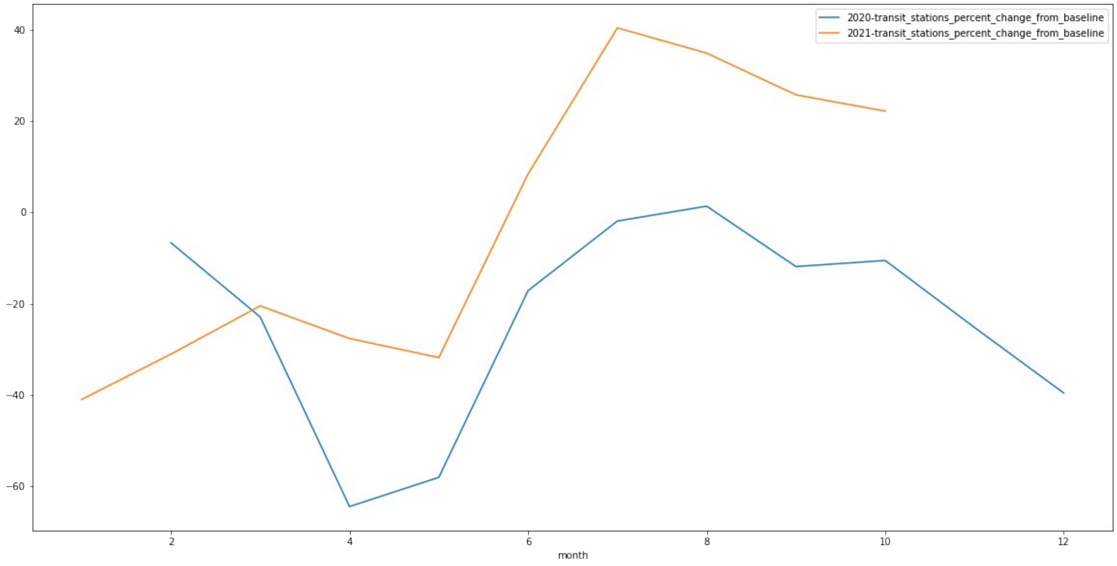
```
null   float64
 3   parks_percent_change_from_baseline                  87099 non-
null   float64
 4   transit_stations_percent_change_from_baseline       78809 non-
null   float64
 5   workplaces_percent_change_from_baseline             154672 non-
null  float64
 6   residential_percent_change_from_baseline            98407 non-
null   float64
dtypes: float64(6), object(1)
memory usage: 8.5+ MB
```

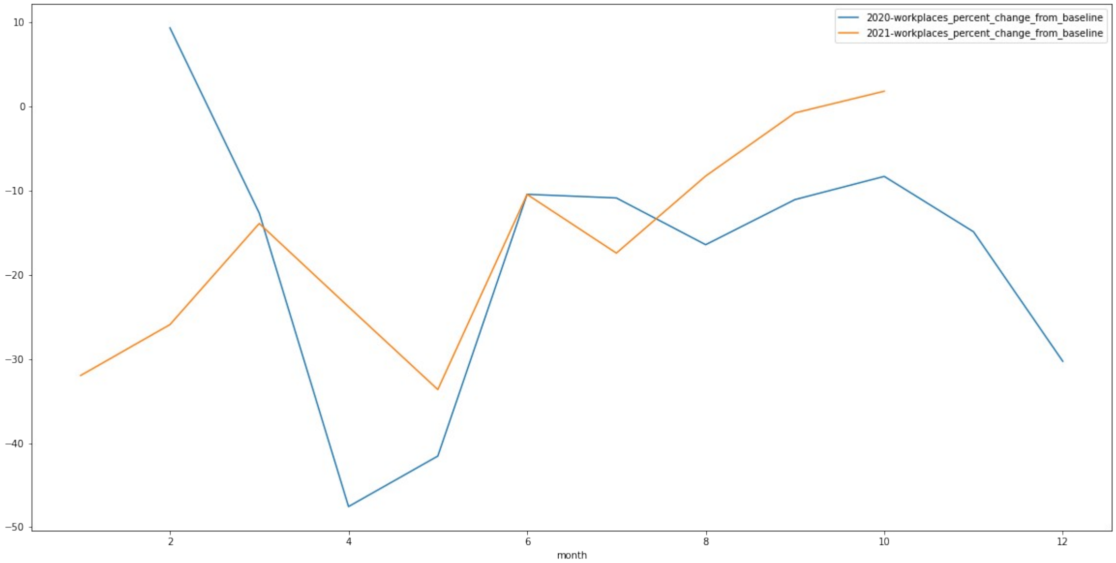
```
<AxesSubplot:xlabel='month'>
```

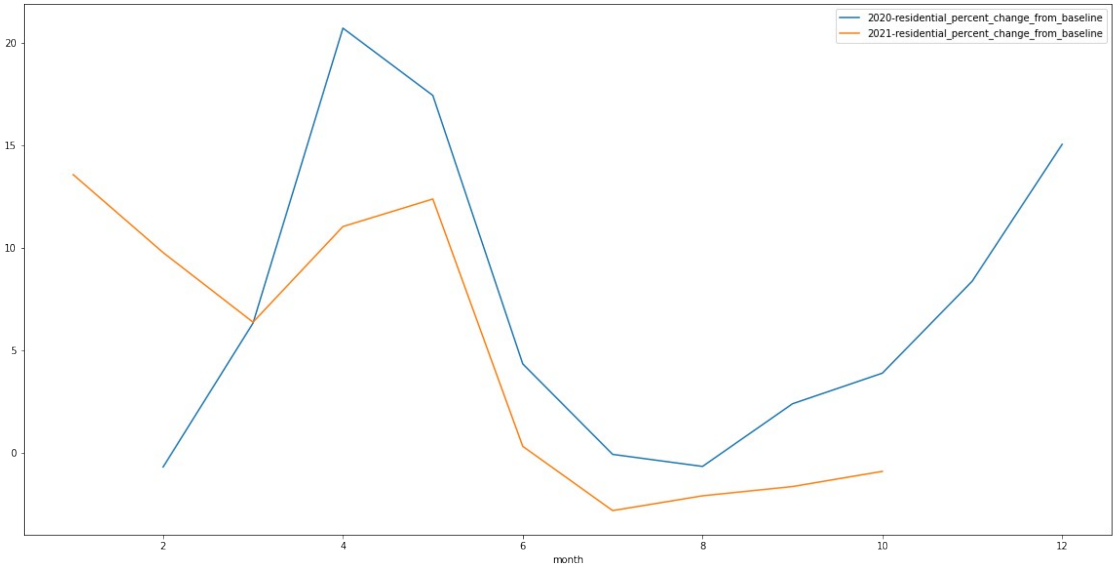





2020-parks_percent_change_from_baseline
2021-parks_percent_change_from_baseline
80
60
40
20
0
−20
−40
2
4
6
8
10
12
month


2020-transit_stations_percent_change_from_baseline
2021-transit_stations_percent_change_from_baseline
40
20
0
−20
−40
−60
2
4
6
8
10
12
month


2020-workplaces_percent_change_from_baseline
2021-workplaces_percent_change_from_baseline
10
0
−10
−20
−30
−40
−50
2
4
6
8
10
12
month


2020-residential_percent_change_from_baseline
2021-residential_percent_change_from_baseline
20
15
10
5
0
2
4
6
8
10
12
month