Unifying Data Science

Difference in Difference for Violent Crime and Marijuana Decriminatlization/Legalization

Exercise 1: Import Data ¶

In [4]:
```
import warnings
warnings.simplefilter('ignore')
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df=pd.read_csv("UDS_arrest_data.csv")
df.head(5)
```

Out[4]:

	YEAR	COUNTY	VIOLENT	F_DRUGOFF	total_population
0	1980	Alameda County	4504	3569	1105379.0
1	1981	Alameda County	4699	3926	1122759.3
2	1982	Alameda County	4389	4436	1140139.6
3	1983	Alameda County	4500	5086	1157519.9
4	1984	Alameda County	3714	5878	1174900.2

Unit - aggregated data per county per year for both violent crimes and drug crimes.

In [5]:
```
import numpy as np
print (f'The data was collected over', np.max(df['YEAR'])-np.min(df['YEA
R'])+1, 'years.')
```

```
The data was collected over 39 years.
```

Exercise 2: Average Drug Rate

In [6]:
```
Drugs09 = df[(df['YEAR'] >= 2007) & (df['YEAR'] <=2009)]
Drugs09.head()
```

Out[6]:

	YEAR	COUNTY	VIOLENT	F_DRUGOFF	total_population
27	2007	Alameda County	4443	6071	1490312.0
28	2008	Alameda County	4336	5893	1496965.0
29	2009	Alameda County	4318	5749	1503618.0
66	2007	Alpine County	8	1	1184.9
67	2008	Alpine County	4	4	1181.6

In [7]:
```
#Calculate the average value of F_DRUGOFF per county.
county_avg=Drugs09.groupby('COUNTY')['F_DRUGOFF'].mean()
#Calculate the median of all average values.
median=Drugs09['F_DRUGOFF'].median()
df['Median']=median
county_avg=pd.DataFrame(county_avg)
county_avg['COUNTY']=county_avg.index
df=df.merge(county_avg, on='COUNTY', how='left')
#Difference between each observation and the global median value.
df['difference']=df['F_DRUGOFF_y']-df['Median']
df['Treated'] = df['difference'].apply(lambda x : 1 if x > 0 else 0)
```

Exercise 3: Determine Violent Crime Rate per 100,000 in population

In [8]:
```
df['Violent_rate'] = df['VIOLENT']/df['total_population']*100000
df.head()
```

Out[8]:

	YEAR	COUNTY	VIOLENT	F_DRUGOFF_x	total_population	Median	F_DRUGOFF_y	differen
0	1980	Alameda County	4504	3569	1105379.0	518.5	5904.333333	5385.8333
1	1981	Alameda County	4699	3926	1122759.3	518.5	5904.333333	5385.8333
2	1982	Alameda County	4389	4436	1140139.6	518.5	5904.333333	5385.8333
3	1983	Alameda County	4500	5086	1157519.9	518.5	5904.333333	5385.8333
4	1984	Alameda County	3714	5878	1174900.2	518.5	5904.333333	5385.8333

Exercise 4: Calculate Difference in Difference by 'hand.'

In [9]:
```
treatedPRE = df[(df['Treated'] == 1) & (df['YEAR'] >=2007) & (df['YEAR']
<=2009)]['Violent_rate'].mean()
controlPRE = df[(df['Treated'] == 0) & (df['YEAR'] >=2007) & (df['YEAR']
<=2009)]['Violent_rate'].mean()
treatedPOST = df[(df['Treated'] == 1) & (df['YEAR'] >=2016) & (df['YEAR'
]<=2018)]['Violent_rate'].mean()
controlPOST= df[(df['Treated'] == 0) & (df['YEAR'] >=2016) & (df['YEAR']
<=2018)]['Violent_rate'].mean()
```

In [10]:
```
effect = (treatedPOST - treatedPRE) - (controlPOST - controlPRE)
print(f'Our effect for the difference in difference pre and post policy
 change is', effect)
```

```
Our effect for the difference in difference pre and post policy change
is -49.40632903015961
```

This tells us that in places of high violent crime, the crime came down by a rate of 49.41 after marijuana decriminalization. This is similar to the initial hypothesis that the illicity of drugs causes violent crime as criminal organizations vie to control of territory.

If we had just conducted a Pre and Post analysis, we would not have seen the same results, as places with low drug rates did not see the same level of decrease in violent crime.

Exercise 5: Calculate Difference in Difference through Regression

In [11]:
```
df['post2010']=df['YEAR'].apply(lambda x : 1 if x >= 2016 else 0)

temp=df[(df['YEAR'] >=2007) & (df['YEAR']<=2009)]
temp2=df[(df['YEAR'] >=2016) & (df['YEAR']<=2018)]
df2 = pd.concat([temp, temp2])

import statsmodels.formula.api as smf
smf.ols('Violent_rate~Treated*post2010+C(COUNTY)', df2).fit().summary()
```

Out[11]:

OLS Regression Results

Dep. Variable:	Violent_rate	**R-squared:**	0.805
Model:	OLS	**Adj. R-squared:**	0.765
Method:	Least Squares	**F-statistic:**	20.12
Date:	Fri, 28 Feb 2020	**Prob (F-statistic):**	4.61e-73
Time:	10:19:23	**Log-Likelihood:**	-1853.2
No. Observations:	348	**AIC:**	3826.
Df Residuals:	288	**BIC:**	4058.
Df Model:	59		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	364.6513	5.530	65.938	0.000	353.767	375.536
C(COUNTY)[T.Alpine County]	101.0447	21.940	4.606	0.000	57.862	144.228
C(COUNTY)[T.Amador County]	-143.9426	21.940	-6.561	0.000	-187.125	-100.760
C(COUNTY)[T.Butte County]	40.7174	31.555	1.290	0.198	-21.390	102.824
C(COUNTY)[T.Calaveras County]	40.8835	21.940	1.863	0.063	-2.299	84.066
C(COUNTY)[T.Colusa County]	-23.9543	21.940	-1.092	0.276	-67.137	19.229
C(COUNTY)[T.Contra Costa County]	-13.6072	31.555	-0.431	0.667	-75.714	48.500
C(COUNTY)[T.Del Norte County]	192.8918	21.940	8.792	0.000	149.709	236.075
C(COUNTY)[T.El Dorado County]	-75.0912	21.940	-3.423	0.001	-118.274	-31.908
C(COUNTY)[T.Fresno County]	207.0049	31.555	6.560	0.000	144.898	269.112
C(COUNTY)[T.Glenn County]	23.8476	21.940	1.087	0.278	-19.335	67.030
C(COUNTY)[T.Humboldt County]	86.4667	31.555	2.740	0.007	24.360	148.574
C(COUNTY)[T.Imperial County]	38.2888	31.555	1.213	0.226	-23.818	100.396
C(COUNTY)[T.Inyo County]	118.2736	21.940	5.391	0.000	75.091	161.456
C(COUNTY)[T.Kern County]	198.9596	31.555	6.305	0.000	136.853	261.067
C(COUNTY)[T.Kings County]	-22.8839	21.940	-1.043	0.298	-66.067	20.299
C(COUNTY)[T.Lake County]	128.7858	21.940	5.870	0.000	85.603	171.969
C(COUNTY)[T.Lassen County]	-54.5548	21.940	-2.487	0.013	-97.738	-11.372
C(COUNTY)[T.Los Angeles County]	55.5320	31.555	1.760	0.079	-6.575	117.639
C(COUNTY)[T.Madera County]	-25.6404	21.940	-1.169	0.244	-68.823	17.542
C(COUNTY)[T.Marin County]	-164.5011	21.940	-7.498	0.000	-207.684	-121.318
C(COUNTY)[T.Mariposa County]	-35.0707	21.940	-1.598	0.111	-78.254	8.112
C(COUNTY)[T.Mendocino County]	189.1406	31.555	5.994	0.000	127.034	251.247
C(COUNTY)[T.Merced County]	175.2687	31.555	5.554	0.000	113.162	237.376

C(COUNTY)[T.Modoc County]	179.5909	21.940	8.186	0.000	136.408	222.774
C(COUNTY)[T.Mono County]	4.1196	21.940	0.188	0.851	-39.063	47.303
C(COUNTY)[T.Monterey County]	51.7667	31.555	1.641	0.102	-10.340	113.874
C(COUNTY)[T.Napa County]	-97.6477	21.940	-4.451	0.000	-140.831	-54.465
C(COUNTY)[T.Nevada County]	-150.8692	21.940	-6.876	0.000	-194.052	-107.686
C(COUNTY)[T.Orange County]	-70.1970	31.555	-2.225	0.027	-132.304	-8.090
C(COUNTY)[T.Placer County]	-52.2457	31.555	-1.656	0.099	-114.353	9.861
C(COUNTY)[T.Plumas County]	106.4001	21.940	4.850	0.000	63.217	149.583
C(COUNTY)[T.Riverside County]	-3.8473	31.555	-0.122	0.903	-65.954	58.260
C(COUNTY)[T.Sacramento County]	110.8686	31.555	3.514	0.001	48.762	172.976
C(COUNTY)[T.San Benito County]	15.4099	21.940	0.702	0.483	-27.773	58.593
C(COUNTY)[T.San Bernardino County]	168.1644	31.555	5.329	0.000	106.057	230.271
C(COUNTY)[T.San Diego County]	30.6849	31.555	0.972	0.332	-31.422	92.792
C(COUNTY)[T.San Francisco County]	84.2993	31.555	2.672	0.008	22.192	146.406
C(COUNTY)[T.San Joaquin County]	187.7026	31.555	5.948	0.000	125.596	249.810
C(COUNTY)[T.San Luis Obispo County]	-129.5777	21.940	-5.906	0.000	-172.761	-86.395
C(COUNTY)[T.San Mateo County]	-82.0348	31.555	-2.600	0.010	-144.142	-19.928
C(COUNTY)[T.Santa Barbara County]	28.8060	31.555	0.913	0.362	-33.301	90.913
C(COUNTY)[T.Santa Clara County]	-30.6416	31.555	-0.971	0.332	-92.749	31.465
C(COUNTY)[T.Santa Cruz County]	9.8490	31.555	0.312	0.755	-52.258	71.956
C(COUNTY)[T.Shasta County]	-78.2402	21.940	-3.566	0.000	-121.423	-35.057
C(COUNTY)[T.Sierra County]	103.5397	21.940	4.719	0.000	60.357	146.723
C(COUNTY)[T.Siskiyou County]	30.8033	21.940	1.404	0.161	-12.380	73.986
C(COUNTY)[T.Solano County]	128.8165	31.555	4.082	0.000	66.710	190.923
C(COUNTY)[T.Sonoma County]	21.3606	31.555	0.677	0.499	-40.746	83.468
C(COUNTY)[T.Stanislaus County]	186.3526	31.555	5.906	0.000	124.246	248.460
C(COUNTY)[T.Sutter County]	100.5170	21.940	4.581	0.000	57.334	143.700
C(COUNTY)[T.Tehama County]	1.6770	21.940	0.076	0.939	-41.506	44.860
C(COUNTY)[T.Trinity County]	122.8432	21.940	5.599	0.000	79.660	166.026
C(COUNTY)[T.Tulare County]	232.5847	31.555	7.371	0.000	170.478	294.692
C(COUNTY)[T.Tuolumne County]	-56.4772	21.940	-2.574	0.011	-99.660	-13.294
C(COUNTY)[T.Ventura County]	-0.6693	31.555	-0.021	0.983	-62.776	61.438
C(COUNTY)[T.Yolo County]	22.1551	31.555	0.702	0.483	-39.952	84.262
C(COUNTY)[T.Yuba County]	219.2809	21.940	9.995	0.000	176.098	262.464
Treated	-66.8063	22.636	-2.951	0.003	-111.359	-22.254
post2010	1.6125	8.287	0.195	0.846	-14.698	17.923

Treated:post2010	-49.4063	11.719	-4.216	0.000	-72.472	-26.340

Omnibus:	39.183	**Durbin-Watson:**	1.751
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	215.714
Skew:	0.187	**Prob(JB):**	1.44e-47
Kurtosis:	6.839	**Cond. No.**	9.97e+15

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 6.06e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Our estimate is the same as when conducted manually as in exercise 4. This tells us that the interaction between two indicator variables is the same as doing the subraction of means from pre and post, from both control and treatment groups.

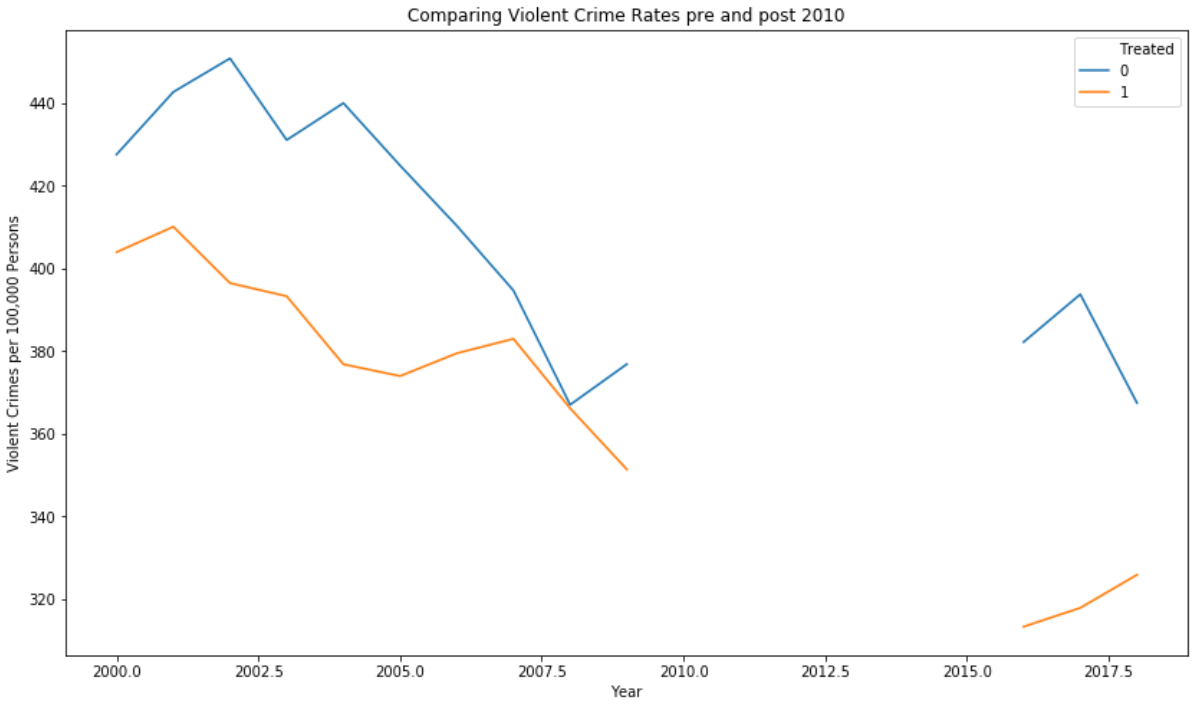
Plot the values to see trends

In [12]:
```
temp=df[(df['YEAR'] >=2000) & (df['YEAR']<=2009)]
temp2=df[(df['YEAR'] >=2016) & (df['YEAR']<=2018)]
```

In [13]:
```
plt.figure(figsize = (14,8))
plt.title('Comparing Violent Crime Rates pre and post 2010')
ax = sns.lineplot(x="YEAR", y="Violent_rate", data=temp, hue='Treated',ci=None)
ax2=sns.lineplot(x="YEAR", y="Violent_rate", data=temp2, hue='Treated',ci=None, legend=False)

plt.xlabel('Year')
plt.ylabel('Violent Crimes per 100,000 Persons')
```

Out[13]: Text(0, 0.5, 'Violent Crimes per 100,000 Persons')



In the prior dataset looking back from 2000 through 2009, we see that both treated and control groups see an overall downward trend, despite some variance year over year. From this, we have more confidence that our Diff-in-Diff is correct as having similar trends is key to the overall validity of our process.

Exercise 7

In [14]:
```
from linearmodels import PanelOLS
df = df.set_index(['COUNTY','YEAR'])
model=PanelOLS.from_formula('Violent_rate ~ Treated * post2010 + EntityEffects', data = df,
                            drop_absorbed=True)
model.fit(cov_type = 'clustered', cluster_entity = True)
```

Out[14]:

PanelOLS Estimation Summary

Dep. Variable:	Violent_rate	**R-squared:**	0.0089
Estimator:	PanelOLS	**R-squared (Between):**	-0.0093
No. Observations:	2262	**R-squared (Within):**	0.0089
Date:	Fri, Feb 28 2020	**R-squared (Overall):**	-0.0080
Time:	10:19:39	**Log-likelihood**	-1.374e+04
Cov. Estimator:	Clustered		
		F-statistic:	9.8710
Entities:	58	**P-value**	0.0001
Avg Obs:	39.000	**Distribution:**	F(2,2202)
Min Obs:	39.000		
Max Obs:	39.000	**F-statistic (robust):**	9.7345
		P-value	0.0001
Time periods:	39	**Distribution:**	F(2,2202)
Avg Obs:	58.000		
Min Obs:	58.000		
Max Obs:	58.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
post2010	4.0853	12.747	0.3205	0.7486	-20.912	29.083
Treated:post2010	-56.881	17.505	-3.2495	0.0012	-91.209	-22.554

F-test for Poolability: 35.481
P-value: 0.0000
Distribution: F(57,2202)

Included effects: Entity
id: 0x12d19fa58

Even when counting for fixed effects from County and Year, we see that the decrease is identical to our previous method without fixed effect. Therefore we can say that marijuana decriminization caused a decrease in violent crime in counties that had higher drug arrest in comparison to the state average.