DSCI 510 HW5 Final Project Description

Name: Wenyu Yang

USC ID: 2609125538

Proposal:
The topic of my final project is finding the relationship between the percentage of at least one dose, percentage of fully vaccinated, percent of a booster dose, and confirm cases percentage by CA counties. The vaccinations do help decrease the confirmed cases.

Motivation:
The main motivation for me to choose this topic is to provide vaccinations to help decrease the confirmed cases and see which factors receiving 1 dose, completion, and booster will have larger influences on confirmed cases in each CA county. Since not everyone has received the same doses of vaccine, I hope the result of this project may help people to consider the importance of vaccinations and whether to receive booster in the future.

Data Source:
There are three datasets.

1. The counties dataset was web scrapping from Wiki. The link is https://en.wikipedia.org/wiki/List_of_counties_in_California. I scrapped the 58 county names populations and areas and stored them in a single csv file.

	county	population	area
0	Alameda Cou	1,648,556	738
1	Alpine Count	1,235	739
2	Amador Cour	41,259	606
3	Butte County	208,309	1,640
4	Calaveras Co	46,221	1,020
5	Colusa Count	21,917	1,151
6	Contra Costa	1,161,413	720
7	Del Norte Co	28,100	1,008
8	El Dorado Co	193,221	1,712

2. The community_level dataset was collected by an External public API (https://data.cdc.gov/resource/3nnm-4jni.json). I extracted count, date, community level, and case_per_100k from the lastest three weeks and stored them in a single csv file.

	county	date	level	case
0	Alameda County	2/24/22	Low	9.1
1	Alameda County	3/3/22	Medium	7.4
2	Alameda County	3/10/22	Low	6
3	Alameda County	3/24/22	Low	3.8
4	Alameda County	3/17/22	Low	4.1
5	Alameda County	3/31/22	Low	3.1
6	Alameda County	4/7/22	Low	2.3
7	Alpine County	2/24/22	Medium	14.7
8	Alpine County	3/3/22	Low	5.6
9	Alpine County	3/10/22	Low	6.2
10	Alpine County	3/24/22	Low	2.2

3. The vaccinations dataset was collected by External public API(https://data.cdc.gov/resource/8xkx-amqh.json). I extracted county, date, dose_1_numb, completion_number, and booster_number from the lastest 50 days and stored them in a single csv file. For both of the External public API datasets, I used county_list which I extracted from web scrapping.

	county	date	dose_1_num	completion_nu	booster_number	population
0	Alameda County	5/9/22	1519822	1374455	848008	1671329
1	Alameda County	5/8/22	1519637	1374312	847627	1671329
2	Alameda County	5/7/22	1519216	1373985	846803	1671329
3	Alameda County	5/6/22	1518927	1373734	846118	1671329
4	Alameda County	5/5/22	1518566	1373471	845552	1671329
5	Alameda County	5/4/22	1518265	1373215	844965	1671329
6	Alameda County	5/3/22	1518028	1373025	844432	1671329
7	Alameda County	5/2/22	1517925	1372960	844223	1671329
8	Alameda County	5/1/22	1517780	1372837	843824	1671329
9	Alameda County	4/30/22	1517368	1372492	842862	1671329
10	Alameda County	4/29/22	1516833	1372132	841914	1671329

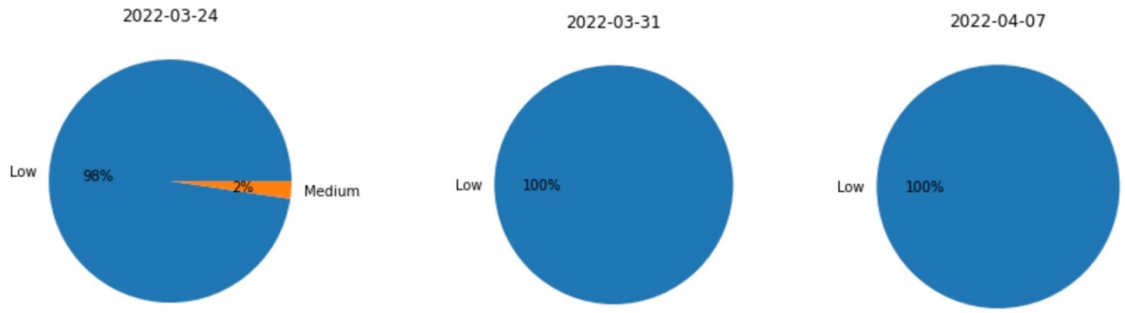
How does the whole combined data system work:

Each dataset contains a list of county names (primary key). The counties dataset has the primary key county, the vaccinations dataset also has the county variable, date, dose_1_numb,completion_number, and booster_number from the lastest 50 days. The community_level dataset also has the county variable, date, level, and case_per_100k. I first contacted three data frames with their common values based on date and county. Then I found out there are empty values (Na) in the data frame, I chose to drop those rows containing NaN and used the rest of the rows to analyze.

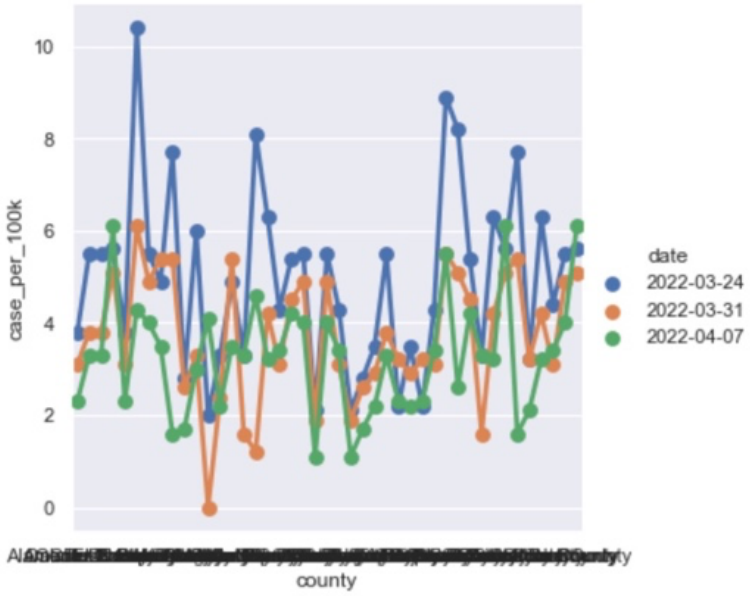
For fitting a multiple regression model and single regression models, I only keep the dose_1_numb,completion_number,booster_number, and case to fit models and draw conclusions.

	dose_1_numb	completion_number	booster_number	case
0	1506492.0	1363793.0	819805.0	380000.0
1	1508616.0	1365623.0	824169.0	310000.0
2	1510429.0	1367237.0	828698.0	230000.0
3	26570.0	20998.0	9497.0	550000.0
4	26637.0	21029.0	9546.0	380000.0
...	...	...	...	...
124	171742.0	154039.0	87382.0	490000.0
125	171948.0	154258.0	87887.0	400000.0
126	47474.0	42374.0	15437.0	560000.0
127	47539.0	42463.0	15584.0	510000.0
128	47629.0	42537.0	15753.0	610000.0

Analysis performed:
I create **three pie charts** based on three days of data I collected. Among those three pie charts, I can tell as time goes on, the number of low levels increases. On 2022-03-24, medium level = 2%, low level=98%. On 2022-03-31, medium level=0%, low level=100%. On 2022-04-07, medium level=0%, low level=100%.The severity of Covid-19 is decreasing.

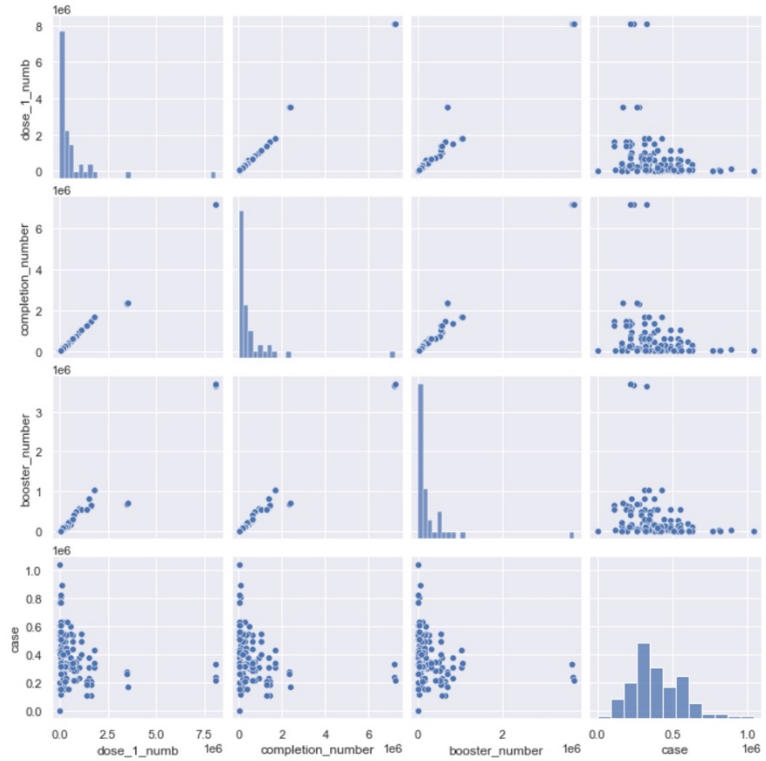


I create **a line plot** using the confirmed cases and county as y and x and categorize them by three different dates. Based on the plot shown, We can conclude that as time goes on, overall confirmed cases are decreasing.



I create **a pairwise plot** between independent variables dose1_num, completion_number, booster_number, and dependent variable case. As the pairwise plot shows, I found that there are

some positive linear relationships between variables. However, there are not enough data to support the conclusion, I need further modeling.



First, I fit **a multiple regression model with all variables** dose_1_numb,completion_number,booster_number. All the p-values are extremely large larger than 0.05. I fail to reject the null hypothesis and indicate there is a multiple linear relationships between dose_1_numb,completion_number,booster_number, and case.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   case   R-squared:                       0.082
Model:                            OLS   Adj. R-squared:                  0.060
Method:                 Least Squares   F-statistic:                     3.738
Date:                Tue, 10 May 2022   Prob (F-statistic):             0.0130
Time:                        00:53:27   Log-Likelihood:                -1731.0
No. Observations:                 129   AIC:                             3470.
Df Residuals:                     125   BIC:                             3481.
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                  4.216e+05   1.63e+04     25.858      0.000    3.89e+05    4.54e+05
dose_1_numb              -0.1693      0.223     -0.758      0.450      -0.611       0.273
completion_number         0.2143      0.401      0.534      0.594      -0.580       1.008
booster_number           -0.1222      0.318     -0.385      0.701      -0.751       0.506
==============================================================================
Omnibus:                       13.237   Durbin-Watson:                   1.588
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.485
Skew:                           0.607   Prob(JB):                     0.000263
Kurtosis:                       4.262   Cond. No.                     2.30e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.3e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Second, I fit **a single regression model between dose_1_number and case**. The p-values= 0.001 which is smaller than 0.05. It indicates there is a linear relationship between dose_1_numb and case. However, the R-squared= 0.079 indicates the accuracy of this model = 7.9% which is extremely low. I need further re-build this model to increase accuracy.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    case   R-squared:                       0.079
Model:                             OLS   Adj. R-squared:                  0.072
Method:                  Least Squares   F-statistic:                     10.92
Date:                 Tue, 10 May 2022   Prob (F-statistic):            0.00124
Time:                         00:53:30   Log-Likelihood:                -1731.2
No. Observations:                  129   AIC:                             3466.
Df Residuals:                      127   BIC:                             3472.
Df Model:                            1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        4.211e+05   1.62e+04     26.058      0.000    3.89e+05    4.53e+05
dose_1_numb    -0.0362      0.011     -3.305      0.001      -0.058      -0.015
==============================================================================
Omnibus:                       13.713   Durbin-Watson:                   1.580
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.803
Skew:                           0.636   Prob(JB):                     0.000225
Kurtosis:                       4.227   Cond. No.                     1.65e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.65e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Third, I fit **a single regression model between completion_number and case**. The p-values= 0.002 which is smaller than 0.05. It indicates there is a linear relationship between completion_number and case. However, the R-squared= 0.076 which indicates the accuracy of this model = 7.6% which is extremely low. I need further re-build this model to increase accuracy.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    case   R-squared:                       0.076
Model:                             OLS   Adj. R-squared:                  0.069
Method:                  Least Squares   F-statistic:                     10.44
Date:                 Tue, 10 May 2022   Prob (F-statistic):            0.00157
Time:                         00:53:32   Log-Likelihood:                -1731.4
No. Observations:                  129   AIC:                             3467.
Df Residuals:                      127   BIC:                             3473.
Df Model:                            1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.206e+05   1.62e+04     25.997      0.000    3.89e+05    4.53e+05
completion_number    -0.0408      0.013     -3.231      0.002      -0.066      -0.016
==============================================================================
Omnibus:                       13.716   Durbin-Watson:                   1.570
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.681
Skew:                           0.641   Prob(JB):                     0.000239
Kurtosis:                       4.209   Cond. No.                     1.43e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.43e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fourth, I fit **a single regression model between booster_number and case**. The p-values= 0.002 which is smaller than 0.05. It indicates there is a linear relationship between booster_number and case. However, the R-squared= 0.070 indicates the accuracy of this model = 7.0% which is extremely low. I need further re-build this model to increase accuracy.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   case   R-squared:                       0.070
Model:                            OLS   Adj. R-squared:                  0.063
Method:                 Least Squares   F-statistic:                     9.584
Date:                Tue, 10 May 2022   Prob (F-statistic):            0.00242
Time:                        00:53:34   Log-Likelihood:                -1731.8
No. Observations:                 129   AIC:                             3468.
Df Residuals:                     127   BIC:                             3473.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.194e+05   1.62e+04     25.907      0.000    3.87e+05    4.51e+05
booster_number    -0.0777      0.025     -3.096      0.002      -0.127      -0.028
==============================================================================
Omnibus:                       13.381   Durbin-Watson:                   1.551
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.138
Skew:                           0.631   Prob(JB):                     0.000313
Kurtosis:                       4.187   Cond. No.                     7.18e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.18e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Describe any technical challenges for solving the problem and how you overcame them:
There are two main technical challenges I faced.

1. When plotting a pairwise plot, only a histogram shows between case and case variables. Rest plots of the variables did not show.
2. There are Na values (empty values) in the data frame. To avoid them during modeling. I need to drop all the rows that contain empty values.
3. I need to call the function I wrote in other .py files when extracting data from external API. For example, the get_community_level function in community_level_api.py needs a list of the county as the query. The list of the county can be obtained from get_county() in counties_web_scraping.py.

Overcame:

1. I googled the seaborn official documents and realized maybe the dtypes for variables are not float. I use type() function to check dtype for each variables, and found out that dose_1_numb,completion_number,booster_number are dtype: object. Then, I used. astype( ) function changed to dtype: float64. Finally, the pairwise plot shows all relationships among all variables.
2. I googled the pandas official documents to solve the problem. First, I used. replace( ) function to replace all ‘Na’ values into np.nan. Then, I used. dropna( ) function to drop those rows.
3. I googled and found out that I can import the counties_web_scraping.py as a package to use get_county() in order to obtain the list of county names.

Conclusion:
Overall, I can conclude that there is a negative relationship between dose_1_number and case. There is a negative relationship between completion_number and case. There is a negative relationship between booster_number and case. It means that the more numbers receiving 1 dose vaccination or completion of vaccination or booster, the lower confirm cases in each county. It also means that no matter if you only receive 1 dose of vaccination or completion of vaccination or booster, the chance of confirming Covid-19 will decrease. The improvement for this project is that the data I collected was not enough to find and perfect a regression model. However, the datasets still confirm my proposal for this project.