1.
This project aims to investigate the relationship between temperature and covid cases. In "Decoding the Role of Temperature in RNA Virus Infections", Bisht indicates that RNA viruses encounter various environments when they copy themselves and spread from host to host and cell to cell. Temperature is one of the factors that affect their stability and transmission. Therefore, this research paper is interested in determining how temperature affects the number of daily covid cases, which implies the viruses' transmissibility. The research sample data is related to the Coronavirus disease 2019 (COVID-19), which is a "highly contagious viral illness caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)", and it is an RNA virus, according to "Features, Evaluation, and Treatment of Coronavirus (COVID-19)". This epidemic has caused more than 6 million deaths worldwide, influencing economics and raising the global health crisis. By finding the relationship between temperature and covid cases, this research can provide suggestions on control temperature factors to decline viruses' activity, stability, and transmissibility to protect human beings.

2.
a)
matplotlib==3.6.2
numpy==1.23.3
pandas==1.5.0
plotly==5.11.0
requests==2.27.1
scipy==1.9.3
seaborn==0.12.1
statsmodels==0.13.5
Also, this information is in my requirements.txt in GitHub.

b)
Running the final.py could re-produce my result. The python script final.py included all the codes used in this research paper. Type "python final.py" in terminal could run it and show the analysis and visualization results.

c)
https://github.com/yufeifeiqiqi/DSCI510_Final_Project

3.
a)
I collected two datasets: historical temperature data and the statewide covid-19 case death tests.

historical temperature data: https://rapidapi.com/visual-crossing-corporation-visual-crossing-corporation-default/api/visual-crossing-weather
Statewide COVID-19 Cases Deaths Tests: https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state/resource/30331e8f-4679-4ee9-908b-df4512065563

API helps collect the temperature data. I will collect all the 2021 daily temperature data information in Los Angeles. Then, I use the date, location of Angeles, lowest temperature, highest temperature, dew point, and relative humidity data.

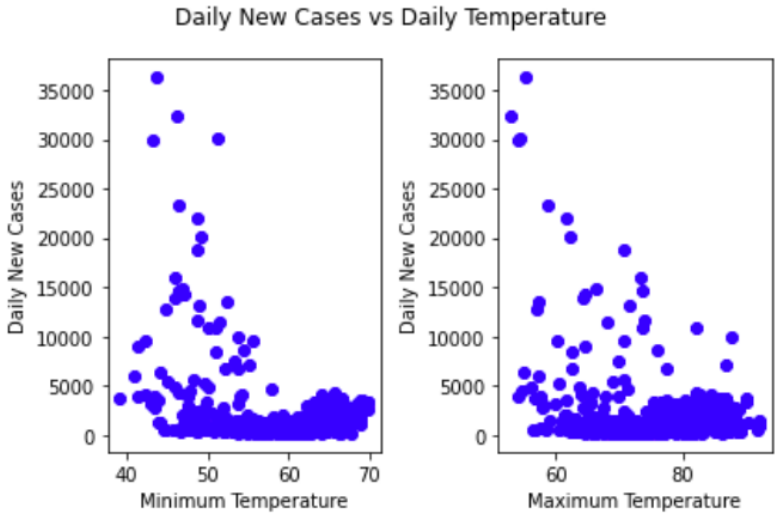
I downloaded a CSV file of the covid case tests dataset from its website. Then, I filtered it by location and date in the Numbers software on Mac to find the index labels of Los Angeles' 2021 dataset. After that, I could get the data I wanted by directly selecting them with the index in python and storing it in a CSV file. I use the date, daily new cases, and deaths information in my analysis. Since my focus is the 2021 daily data, the sample size is 365.

b)
I first tried to use API to collect data from the covid cases tests dataset and had a hard time achieving my goal of collecting them. I found out later that it was easier to download it and organize what I needed. As for the temperature datasets, when I first managed them from API, it produced the complete information as a string having commas separating each value. I was surprised to find out why it was not like datasets or data frames, and I asked the teaching assistant for help. I then knew that my collected data was already in the data frame's format, but I needed to create a CSV file and write them into it.

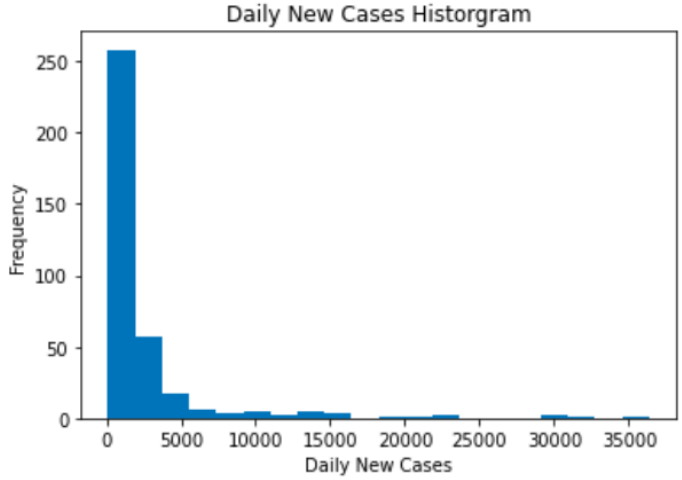
Moreover, this API only allows me to collect continuous data for up to 31 days. Therefore, I needed to collect data twelve times instead of simultaneously managing the whole year's data. I collected twelve datasets into twelve CSV files and merged them into one CSV file.
By selecting what I needed from these temperature and covid cases datasets, I merged them into one for further analysis.

4.
a)
I plotted the scatter plots, a histogram, and a boxplot to see the trend and distribution of the data. Then, I draw polynomial and exponential regression plots with ordinary least-squared methods to approach the data. After that, I did hypothesis testing on the correlation relationship between maximum temperature and covid cases and between minimum temperature and covid cases, respectively.

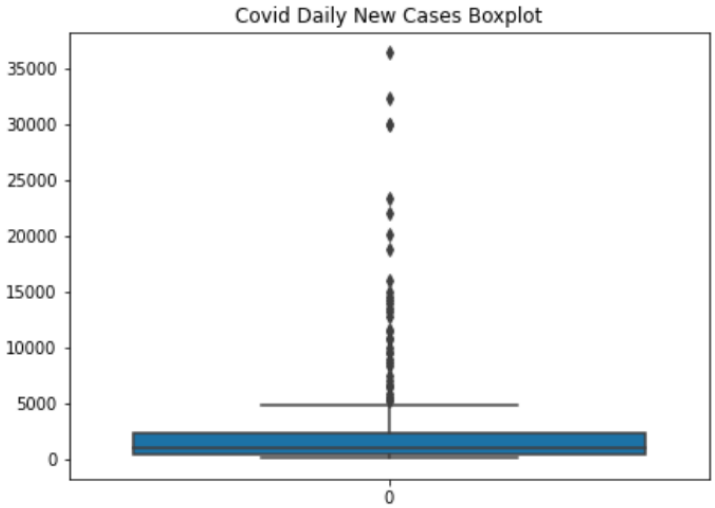
I have done scatter plots for minimum temperature and daily new covid cases and for maximum temperature and daily new covid cases. The figure shows a similar trend; most data are clustering at the bottom, under 5000 cases. But the graph of the minimum temperature and daily new cases has more outliers. I will not drop these outliers from the dataset because I am interested in a year's data information.



I then did a histogram of the daily new covid cases. The result indicates that data are distributed mainly under 5000 cases. This result also supports the previous figure that most data cluster at the bottom. The distribution is not normal and heavily right-skewed. There are some outliers at the position, around 30000 and 35000 cases.



This boxplot displays the distribution of the covid daily cases. It indicates that the lower half of the data is closely distributed and very close to the minimum value. It also shows that there are many outliers.



I got the correlation coefficient for two temperature variables with the case variable separately. The correlation coefficient for maximum temperature and daily covid cases is 0.274, while it is -0.328 for minimum temperature and daily covid cases. This statistical measurement represents the linear relationship between two variables. Therefore, both temperature variables have a weak linear relationship with the dependent variable. But they might have a robust nonlinear relationship. I used polynomial regression to approach the case data. Though they had a nonlinear relationship, the polynomial regression models for two temperature variables were linear.

```
print(np.corrcoef(f_df["Maximum Temperature"], f_df["cases"]))
```

```
[[ 1.         -0.27367657]
 [-0.27367657  1.        ]]
```

```
print(np.corrcoef(f_df["Minimum Temperature"], f_df["cases"]))
```

```
[[ 1.         -0.32761544]
 [-0.32761544  1.        ]]
```

These are mostly all the outliers' information. They all occur during January and December of 2021. It might relate to the temperature variable or be because of the holiday vacation or many other possible reasons.

```
f_df[f_df["cases"]>5000]
```

Date time	Minimum Temperature	Maximum Temperature	Temperature	cases	Address
01/01/2021	45.1	68.9	56.0	5343.0	Los Angeles,CA,USA
01/02/2021	47.1	64.5	54.9	14211.0	Los Angeles,CA,USA
01/03/2021	51.0	62.6	55.6	8366.0	Los Angeles,CA,USA
01/04/2021	48.8	61.8	54.7	22063.0	Los Angeles,CA,USA
01/05/2021	49.3	62.4	54.5	20139.0	Los Angeles,CA,USA
01/06/2021	48.7	70.7	57.6	18737.0	Los Angeles,CA,USA
01/07/2021	46.0	73.5	56.1	15962.0	Los Angeles,CA,USA
01/08/2021	47.0	66.5	56.4	14922.0	Los Angeles,CA,USA
01/09/2021	50.1	73.8	59.7	10814.0	Los Angeles,CA,USA
01/10/2021	48.2	70.9	58.4	5580.0	Los Angeles,CA,USA
01/11/2021	46.4	73.6	58.4	14615.0	Los Angeles,CA,USA
01/12/2021	48.9	71.5	58.7	13228.0	Los Angeles,CA,USA
01/13/2021	48.7	74.1	59.7	11664.0	Los Angeles,CA,USA
01/14/2021	51.1	82.1	64.1	10794.0	Los Angeles,CA,USA
01/15/2021	53.7	87.6	68.3	10028.0	Los Angeles,CA,USA
01/16/2021	55.1	86.8	67.9	7051.0	Los Angeles,CA,USA
01/18/2021	53.8	77.6	63.3	6858.0	Los Angeles,CA,USA
01/19/2021	55.7	70.9	61.1	9508.0	Los Angeles,CA,USA
01/20/2021	54.4	76.0	64.6	8677.0	Los Angeles,CA,USA
01/21/2021	53.4	69.8	61.2	7500.0	Los Angeles,CA,USA
01/22/2021	52.2	62.6	56.8	6866.0	Los Angeles,CA,USA
01/25/2021	44.1	54.9	49.8	6418.0	Los Angeles,CA,USA
01/26/2021	41.0	57.4	49.2	5967.0	Los Angeles,CA,USA
01/27/2021	49.7	60.4	53.7	5288.0	Los Angeles,CA,USA
12/20/2021	41.3	64.7	52.4	9008.0	Los Angeles,CA,USA
12/21/2021	51.4	68.1	56.8	11485.0	Los Angeles,CA,USA
12/22/2021	46.0	64.2	56.5	13905.0	Los Angeles,CA,USA
12/23/2021	52.5	57.2	54.5	13558.0	Los Angeles,CA,USA
12/24/2021	42.3	60.2	55.3	9552.0	Los Angeles,CA,USA
12/26/2021	44.8	57.1	51.3	12825.0	Los Angeles,CA,USA
12/27/2021	43.3	54.1	49.4	29872.0	Los Angeles,CA,USA
12/28/2021	43.7	55.1	49.1	36345.0	Los Angeles,CA,USA
12/29/2021	46.2	52.9	49.7	32272.0	Los Angeles,CA,USA
12/30/2021	51.3	54.3	53.0	30017.0	Los Angeles,CA,USA
12/31/2021	46.4	58.9	52.8	23316.0	Los Angeles,CA,USA

I used polynomial and exponential regression with the ordinary least squared method to approach the maximum temperature data in covid cases. Polynomial regression is a good fit, while exponential regression does not fit well. Therefore, I only did the polynomial regression of the minimum temperature variable on covid cases.

This is the coefficients for the polynomial function for maximum temperature.

```
def func(x, a, b, c):
    y = a*x*x + b*x+c
    return y

alpha, beta, intercept = optimize.curve_fit(func, xdata = x, ydata = y)[0]
print(f'alpha={alpha}, beta={beta}, intercept = {intercept}')
```

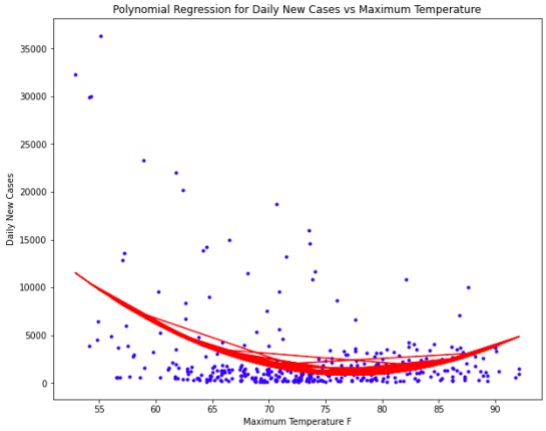
```
alpha=18.135717611263434, beta=-2800.0103960413016, intercept = 108894.94386337163
```

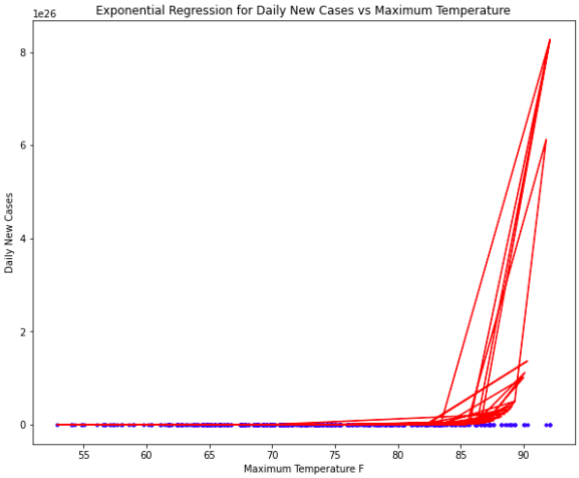
This is the coefficients of parameters for the polynomial function for minimum temperature.

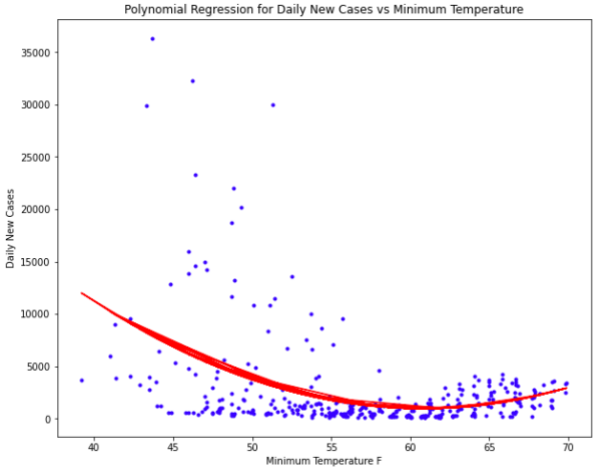
```
def func(x, a, b, c):
    y = a*x*x + b*x+c
    return y

alpha, beta, intercept = optimize.curve_fit(func, xdata = x, ydata = y)[0]
print(f'alpha={alpha}, beta={beta}, intercept = {intercept}')
```

```
alpha=23.68981666653756, beta=-2880.2125452059577, intercept = 88502.63335505218
```







I want to know the relationship between temperature and, if there is a relationship, which temperature variable is more statistically significant. Thus, I did a hypothesis test twice on two independent variables, the maximum and the minimum temperature, separately since they were highly correlated, as seen from the scatter plots of their trend.

Although my covid cases data is not normally distributed, my sample size of 365 is large enough that these kinds of traditional null hypothesis significance tests are robust to violations of this assumption.

My null hypothesis is that there is no relationship between temperature and daily covid cases, while my alternative hypothesis is that there is a relationship between temperature and daily covid cases. I did an independent t test for both tests because my independent variables and dependent variables were continuous.

The p-value for the test of the Maximum Temperature variable on the cases variable is 5.495×10^{-21} while the p-value for the test of the Minimum Temperature on the cases is 3.212×10^{-21}. Both p-values are smaller than the 0.05 significant value, which indicates that they are statistically significant. I can reject the null hypothesis and conclude that there is a relationship between temperature variables and covid cases. Those outliers might contribute

a lot to this result. The p-value of the Minimum Temperature hypothesis test is smaller, so Minimum Temperature is more statistically significant towards daily covid cases.

```
ttest,pval = ttest_ind(f_df["Maximum Temperature"],f_df["cases"],equal_var = False, alternative= "two-sided")
print(f"The statistic t value is {ttest} and the p-value is {pval}")
```

```
The statistic t value is -10.007062739291262 and the p-value is 5.4953356644709336e-21
```

```
ttest,pval = ttest_ind(f_df["Minimum Temperature"],f_df["cases"],equal_var = False, alternative= "two-sided")
print(f"The statistic t value is {ttest} and the p-value is {pval}")
```

```
The statistic t value is -10.074685356004688 and the p-value is 3.21249495126667e-21
```

The following figure presents the coefficients of the polynomial regression for the Maximum temperature to fit the cases. The determination is the R-squared value of 0.194. this value is weak because it results from the residuals sum of squares divided by the total sum of squares. Since we have several extreme outliers far from the fitted model and their corresponding predictive value, and when we square the residuals, the distance between the observed and predictive values, we make the distances even more prominent by adding into the residuals of squares. The same thing happened when calculating the R-squared value for the Maximum temperature polynomial regression model, which is 0.181.

```
{'polynomial': [18.1357162379303, -2800.0101940512172, 108894.93653946006], 'determination': 0.19431642161659518}
```

```
{'polynomial': [23.68981727894456, -2880.212615210808, 88502.63532531407], 'determination': 0.1810859969573649}
```

Since R-squared values are very close and do not explain the model much in this case, I used minimum temperature in further analysis.

After finding the relationship, I wanted to add more variables inside in temperature data to form new models. I considered adding the dew point and relative humidity. I created a new data frame that contained all I needed for further analysis. I first calculated the correlation matrix to decide which variables to use. From the following figure, since point and relative humidity are highly correlated (0.724), I only chose one of these variables. In addition, since minimum temperature and dew point are highly correlated (0.788) while minimum temperature and relative humidity are weakly correlated (0.210), I included relative humidity in further modeling.

```
Correlation matrix is :
                    Minimum_Temperature  Dew_Point  Relative_Humidity  \
Minimum_Temperature             1.000000   0.787565           0.209791
Dew_Point                       0.787565   1.000000           0.724369
Relative_Humidity               0.209791   0.724369           1.000000
cases                          -0.327615  -0.168901           0.092554

                         cases
Minimum_Temperature -0.327615
Dew_Point           -0.168901
Relative_Humidity    0.092554
cases                1.000000
```

In my following analyses and my visualizations, “min2” represents the squared value of minimum temperature. “Minimum_Temperature” is the minimum temperature value. “Relative_Humidity” is the relative humidity value. The “min_humidity” is the interaction effect of minimum temperature and relative humidity by multiply them.
This is my previous polynomial model, and it has Adjusted R-squared 0.177.

```
model = ols(formula='cases ~ min2 + Minimum_Temperature   ', data=f2_df).fit()
summary = model.summary()
summary
```

OLS Regression Results

Dep. Variable:	cases	**R-squared:**	0.181
Model:	OLS	**Adj. R-squared:**	0.177
Method:	Least Squares	**F-statistic:**	40.02
Date:	Mon, 12 Dec 2022	**Prob (F-statistic):**	1.98e-16
Time:	17:21:03	**Log-Likelihood:**	-3557.5
No. Observations:	365	**AIC:**	7121.
Df Residuals:	362	**BIC:**	7133.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.85e+04	1.31e+04	6.775	0.000	6.28e+04	1.14e+05
min2	23.6898	4.149	5.710	0.000	15.531	31.849
Minimum_Temperature	-2880.2126	468.157	-6.152	0.000	-3800.862	-1959.563

I added the relative humidity variable into the model, and the Adjusted R-squared value is 0.198. the difference between R-squared and Adjusted R-squared is that R-squared tends to estimate the fit of the linear regression optimistically, and it always increases as the number of effects is included in the model. Adjusted R-squared attempts to correct this overestimation, so it might decrease when adding new variables that do not improve the model. Since the Adjusted R-squared increases, this new model fits the covid cases data better than the previous model.

```
model = ols(formula='cases ~ min2 + Minimum_Temperature + Relative_Humidity   ', data=f2_df).fit()
summary = model.summary()
summary
```

OLS Regression Results

Dep. Variable:	cases	**R-squared:**	0.204
Model:	OLS	**Adj. R-squared:**	0.198
Method:	Least Squares	**F-statistic:**	30.86
Date:	Mon, 12 Dec 2022	**Prob (F-statistic):**	8.78e-18
Time:	17:21:03	**Log-Likelihood:**	-3552.3
No. Observations:	365	**AIC:**	7113.
Df Residuals:	361	**BIC:**	7128.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.441e+04	1.3e+04	6.515	0.000	5.89e+04	1.1e+05
min2	23.0384	4.101	5.618	0.000	14.974	31.103
Minimum_Temperature	-2828.0608	462.445	-6.115	0.000	-3737.484	-1918.637
Relative_Humidity	51.9102	16.057	3.233	0.001	20.333	83.487

As last, I added the interaction effect of the minimum temperature and relative humidity into the model. I got an Adjusted R-squared 0.233, much higher than not adding the interaction

variable.

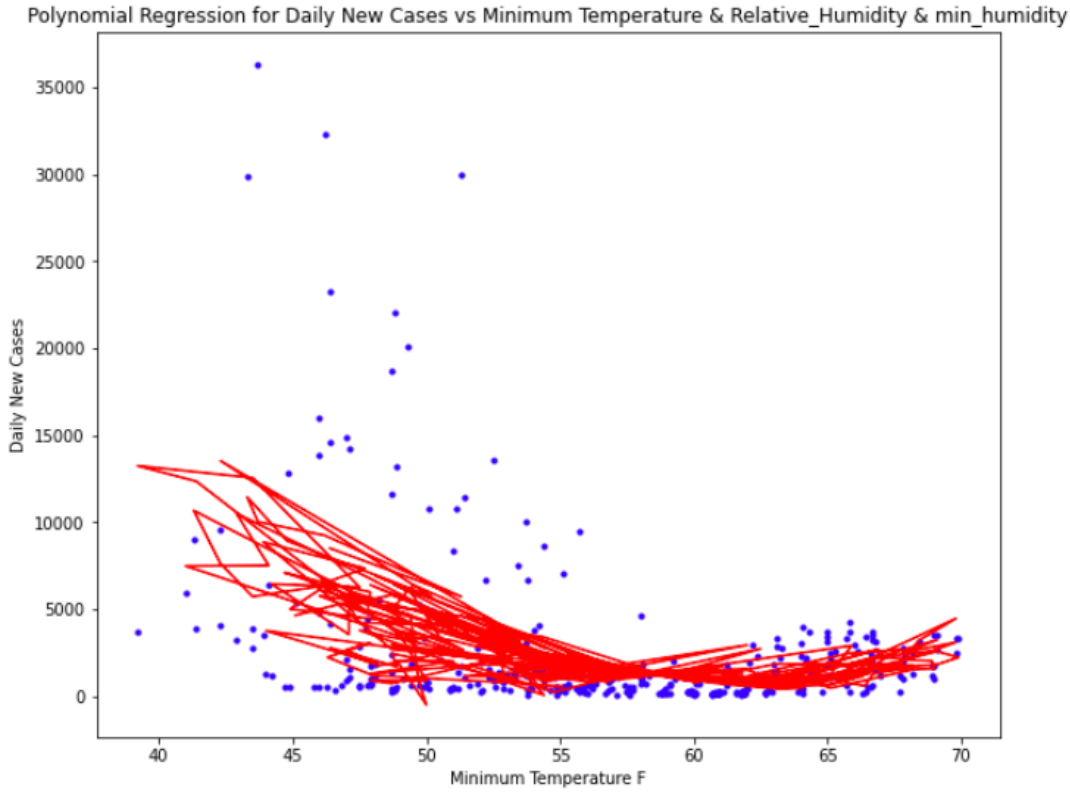
```
model = ols(formula='cases ~ min2 + Minimum_Temperature + Relative_Humidity + min_humidity ', data=f2_df).fit()
summary = model.summary()
summary
```

OLS Regression Results

Dep. Variable:	cases	**R-squared:**	0.241
Model:	OLS	**Adj. R-squared:**	0.233
Method:	Least Squares	**F-statistic:**	28.62
Date:	Mon, 12 Dec 2022	**Prob (F-statistic):**	1.16e-20
Time:	17:21:03	**Log-Likelihood:**	-3543.6
No. Observations:	365	**AIC:**	7097.
Df Residuals:	360	**BIC:**	7117.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.039e+04	1.5e+04	3.350	0.001	2.08e+04	8e+04
min2	25.0587	4.038	6.205	0.000	17.117	33.000
Minimum_Temperature	-2306.0375	468.955	-4.917	0.000	-3228.273	-1383.802
Relative_Humidity	682.8575	151.159	4.517	0.000	385.592	980.123
min_humidity	-11.6713	2.781	-4.197	0.000	-17.140	-6.202

I plotted this new model, but it seems overfitting. After comparing this model with the initial polynomial regression model, only having the minimum temperature variable, it might be better to use the initial polynomial regression model.



b)

I operated hypothesis tests and polynomial regression models and conclude my alternative hypothesis that there is a relationship between temperature variables and covid cases. Moreover, from the p-values I got, minimum temperature variable is more statistically significant on covid cases than maximum temperature. After that, I wanted to add dew point

and relative humidity variables into my model. But after calculating the correlation matrix, I realized that dew point is highly correlated with relative humidity, which means I only need one of them. Furthermore, dew point is also highly correlated with minimum temperature while relative humidity is weakly correlated. As a result, I chose relative humidity into my new model. I first tested the Adjusted R-squared value my original polynomial regression model for minimum temperature. And then, adding relative humidity into that model and got an larger Adjusted R-squared value. I was also interested the interaction effect of minimum temperature and relative humidity on covid cases, so I added the interaction term into the model. I got a larger Adjusted R-squared value. However, after drawing the plot for my final multilinear regression model having interaction variable, the plot seems overfitted to the covid cases data. it would not be a good model if I change the sample. Therefore, the original polynomial regression model with only minimum temperature variable might be a better fit for the covid cases data.

This project fitted models with our temperature data to the covid cases' information to understand more about RNA virus and provide possible further investing questions to reduce virus transmission and stability and protect human beings. Temperature changes relate to every person's life. It leads to further research questions as to whether controlling the inner room's temperature prevents the spread of the virus or if we can find the weakness of the virus if they have to rely on some specific growing environment.

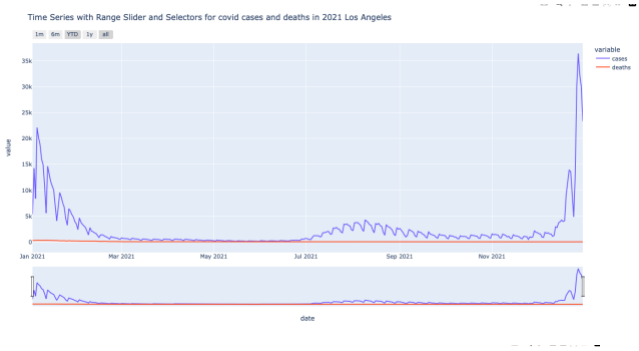
c)

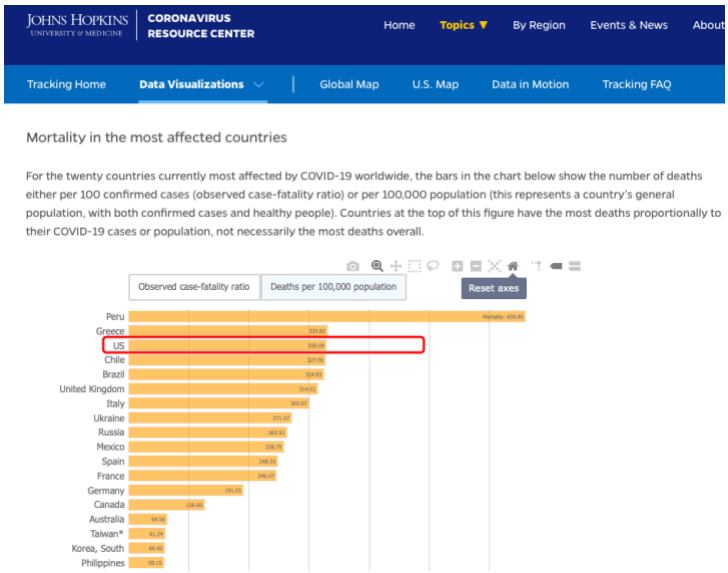
I met a problem when doing factorial ANOVA tests for the model containing minimum temperature, relative humidity, and their interaction effect. The computer kept running and could not produce results when I put all 365 days' data into it to calculate the model's coefficients. After entering monthly 31 days data to see what happened, I realized that the computer thought each unique value in minimum temperature and relative humidity variables was a treatment group. It tried to interact with other groups having different treatments. So it might produce 31 times 31 interaction treatment groups. However, my data did not have any treatments. I researched and figured out I wanted a multilinear regression model with an interactive variable. So, I changed my models and codes, leading to what I wanted to test.

Advanced visualizations

I did two advanced visualizations.

The first visualize daily cases and deaths through 2021 in Los Angeles. From this figure, it is evident that there were explosions of covid in January and December. The deaths' number does not vary much and has a decreasing trend. From this figure, it is possible to calculate the death rate and compare it with the death rate calculated from the total cases and deaths from Johns Hopkins University Medicine's website. If these two death rates are close to each other, it implies that the sample of 2021 covid cases in Los Angeles is a good sample prediction for the United States covid cases. Moreover, I can search for information on the release date of vaccines and question whether the reason for the enormous decrease in covid cases from March to July is that most people had been vaccinated.

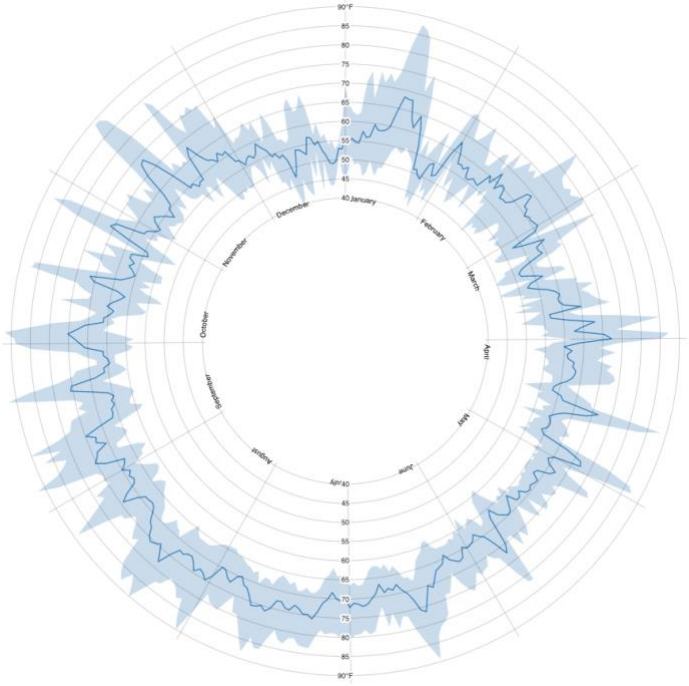




From https://coronavirus.jhu.edu/data/mortality

From https://www.immunize.org/newreleases/viewall2021.asp

The second advanced visualization is about my temperature data written on this website https://observablehq.com/d/f996c24b22e54fb0. This figure could help visualize the trend and changes in our daily temperature in 2021. There is a blue domain for the temperature. The upper bound of this range is the daily maximum temperature, while the lower bound is the daily minimum temperature. From this figure, several days in each month have a really high maximum temperature except in December. Moreover, this figure indicated that Los Angeles is a normally warm city with an extensive daily temperature range.



5. Future work

Since outliers occur in January and December, I would make the data in these two months into a separate model while data in the rest of the year is into another different model. Current models need to fit into the January and December data better. The predictive values giving minimum temperature in January and December are much lower than the observed values. Separating them might help me research factors that relate to covid cases. I am also interested in possible reasons for the vast differences between these two models.

What's more, this project proves that there is a relationship between minimum temperature and covid cases. If I want to investigate further though not in a data science discipline, the next step will be to design an experiment to investigate their causal relationship for future disease prevention and decline in their transmission and break their stability.

Citations

Bisht, Karishma, and Aartjan J. W. Te Velthuis. “Decoding the Role of Temperature in RNA Virus Infections.” MBio, vol. 13, no. 5, 18 Aug. 2022, pp. e02021–22. https://doi.org/10.1128/mbio.02021-22

Cascella M, Rajnik M, Aleem A, et al. Features, Evaluation, and Treatment of Coronavirus (COVID-19) [Updated 2022 Oct 13]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554776/