Lam Tran

AM 170B

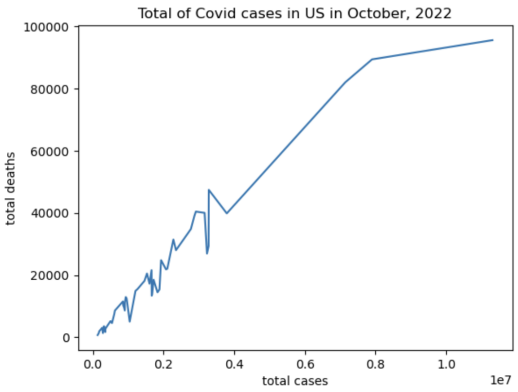
April 15th, 2023

Week 1 Homework: Data exploration and visualization

I-COVID Data Visualization:

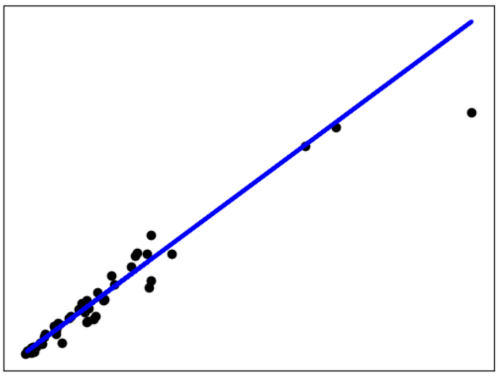
1. Download COVID case and death count dataLinks to an external site. from CDC.gov

2. Plot the total number of cases vs the total number of deaths per state. Do you observe a relationship between these variables?



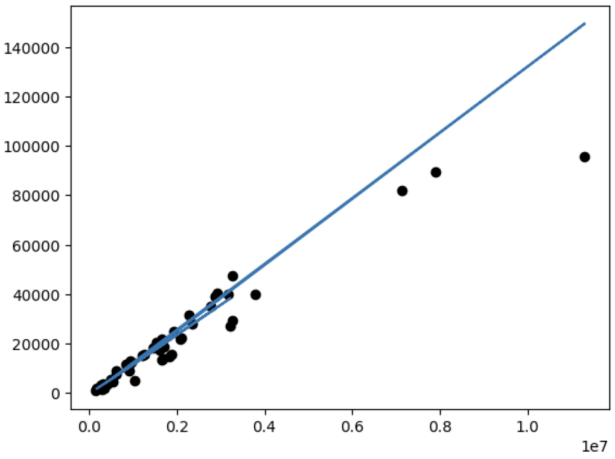
Attempt to fit this data to

a) a linear function



Fit error: 148398542.21

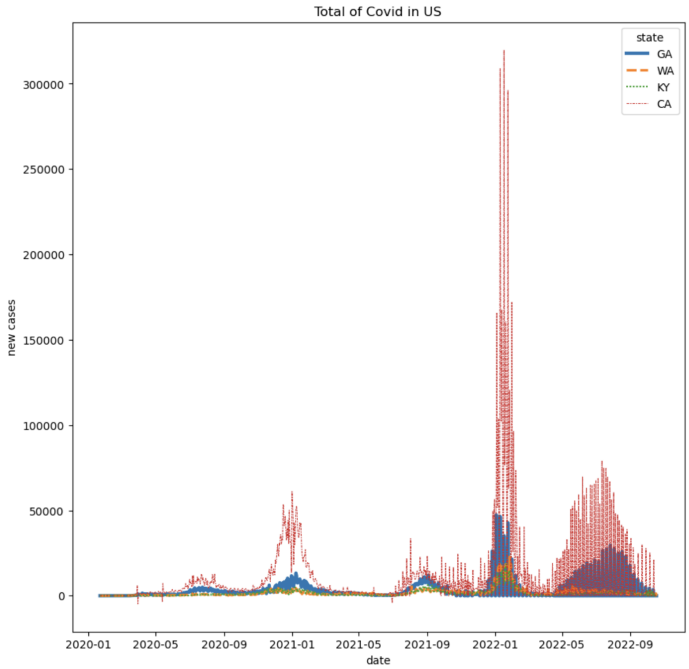
b) a polynomial function.



The graph of the polynomial function doesn't show as a curve because the total cases and the total death are equivalent to increasing. We can conclude the states with a higher number of cases tend to have a higher number of deaths. In conclusion, as The number of cases increases, the number of deaths also increases.

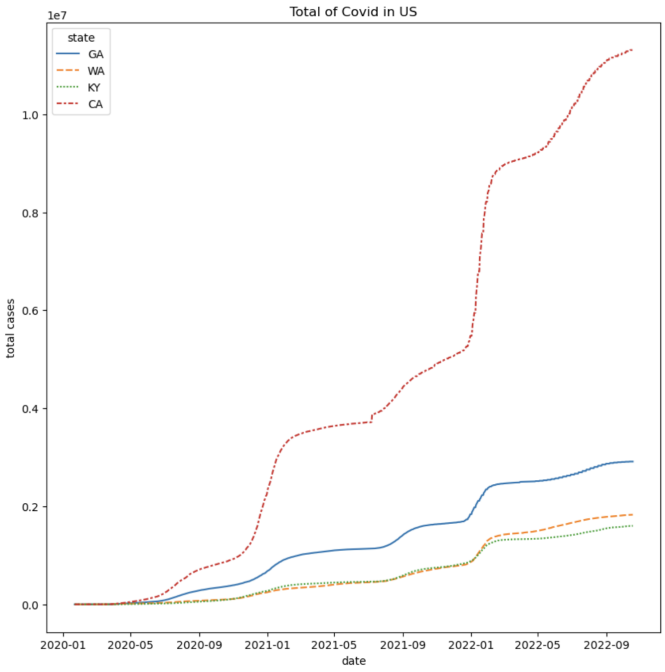
4. Consider data for the following states: Washington, Kentucky, Georgia, and California. And Discuss the results

a) Plot new cases over time



The state that has the newest case is California. California reaches up to 300 000 new cases. The second highest is Georgia. Georgia reaches up to 50 000 new cases. The following state is Washington, Kentucky. The highest peak for all of the states is on January 2022. The lowest is between May 2021 and January 2020.

b) Plot total cases over time.



The new cases for all of the state is increasing by time from 2020 to 2022. The most highest new case is California. California new cases reach up to 10 million case. The second highest is Georgia. Georgia reach up to 2 million cases. The next two states are Kentucky and Washington growth nearly equivalent.

c) Find population count for each state and normalize new cases data and total cases data.

<u>The population of each state in 2021:</u>

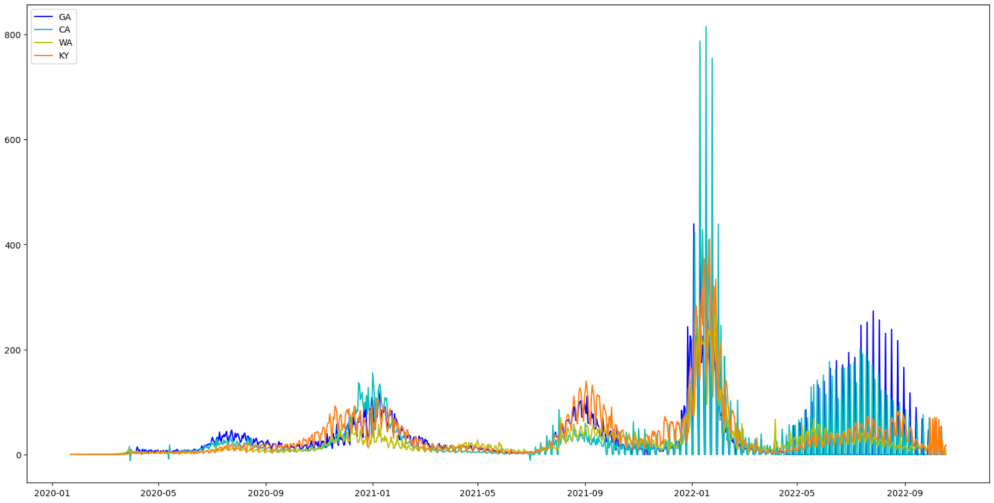
- Washington(WA): 7.739 million
- Kentucky (KY): 4.509 million
- Georgia (GA): 10.8 million
- California(CA): 39.24 million

<u>Normalize new cases data and total cases data:</u>

- (New cases/population size) * 100000
- (total cases/population size) * 100000

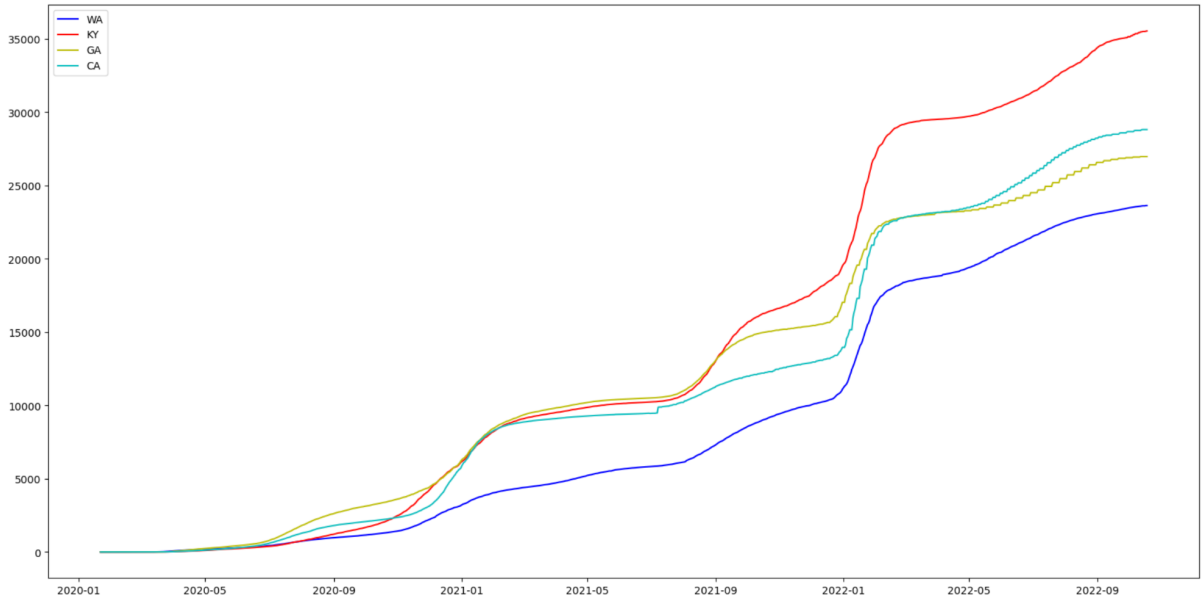
Replot a) and b).

Plot new cases over time



Similarity, California has the highest number of new cases. And growth up to 800 new cases during January 2022. Kentucky and Georgia's new cases grew equivalently.

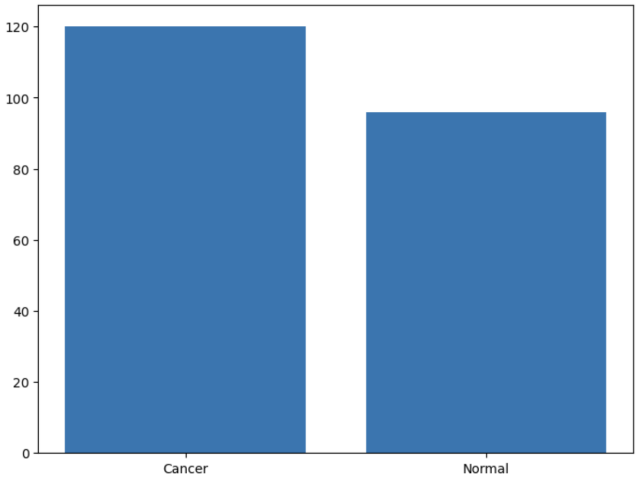
b) Plot total cases over time



After normalizing with the population for each state. The data is changing compared to the previous graph. The highest number of total cases is in Kentucky. Kentucky reaches up to 35 000 cases. California and Georgia grow equivalent and reached up to around 30 000 cases. The least is Washington state reached up to 20 000 cases.

II-Gene expression differences between normal ovary tissue and ovarian cancer biopsies.

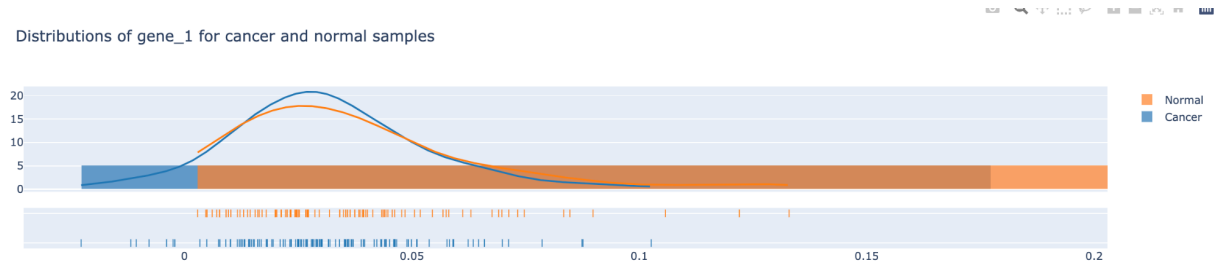
1. Download ovarian_cancer.csv from Canvas. Plot the number of cancer samples and the number of normal samples. How many genes were measured?



Cancer has 120 samples.

Normal has 96 samples

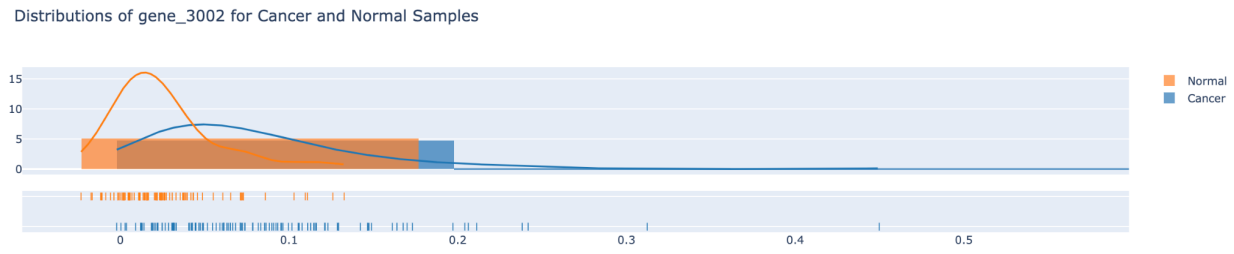
2. Plot distributions of gene_1 for cancer and normal samples.



Calculate the log2 fold change of gene_1 between cancer and normal sample means

the log2 fold : -0.23802293718285011

Plot distributions of gene_3002 for cancer and normal samples.



Calculate the log2 fold change of gene_3002 between cancer and normal sample means

the log2 fold: 1.6255639181537598

a) Use the t-test to determine if there is a significant difference between the means of cancer and normal groups for gene_1 and gene_3002. State the log2 fold changes and the pvalues for each gene.

If the p-value is less than 0.05, then there is a significant difference between the means of cancer and normal groups.

gene_1:

- p-value = 0.08152199430351342

gene_3002:

- p-value = 3.4629258618496455e-12

=> log2 fold changes: 1.88354513

After observing the p-value, there is no significant difference between the means of cancer and normal groups for gene_1 and gene_3002. Since the value is both greater than 0.05

b) Use the Mann-Whitney-Wilcoxon test to determine if there is a significant difference between the means of cancer and normal groups for gene_1 and gene_3002. State the log2 fold changes and the pvalues for each gene.

gene_1:

- p-value = 0.3971031674099791

gene_3002:

- p-value = 2.5643479615401806e-15

=> log2 fold changes: 12.19398729

After observing the p-value, there is no significant difference between the means of cancer and normal groups for gene_1 and gene_3002. Since the value of both cases is greater than 0.05.

c) Discuss the results of these tests. Assume that the significance threshold is when the p-value > 0.05.

The result of p-values for both of the tests is less than 0.05. We can conclude that the null hypothesis is incorrect. There are no significant differences between the means of cancer and normal groups for gene_1 and gene_3002.

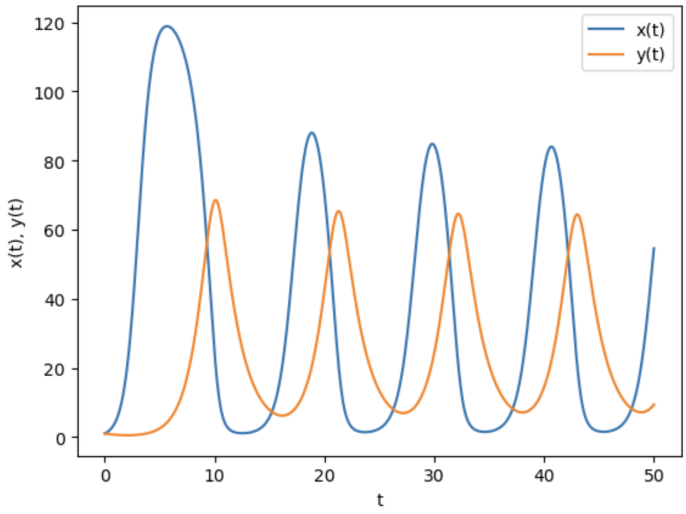
III-Population dynamics: predator-prey.

1. Implement the following population dynamics model:

$\dot{x} = rx(1 - \frac{x}{k}) - \frac{axy}{c+x}, x \geq 0$

$\dot{y} = \frac{baxy}{c+x} - dy, y \geq 0$

The parameter values are a = 3.2, b = 0.6, c = 50, d = 0.56, k = 125, and r = 1.6.



2. Find the three equilibrium points of the system, by setting x'(t) = 0, y'(t) = 0, and solving for x and y. Note one of the equilibrium points is (0,0).

2. Find the three equilibrium points of the system

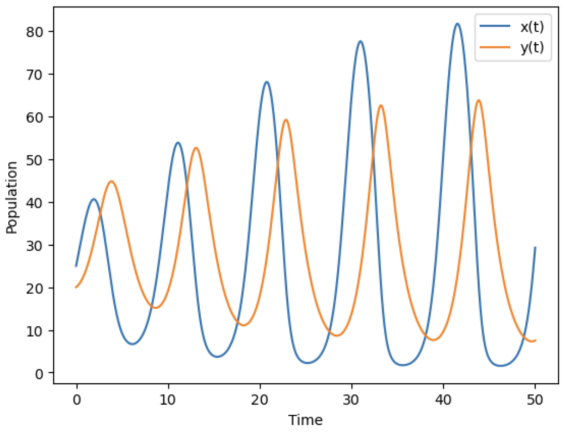
$\dot{x} = rx\left(1 - \frac{x}{k}\right) - \frac{axy}{c+x}, \quad x \geq 0$

$\dot{y} = \frac{baxy}{c+x} - dy, \quad y > 0$

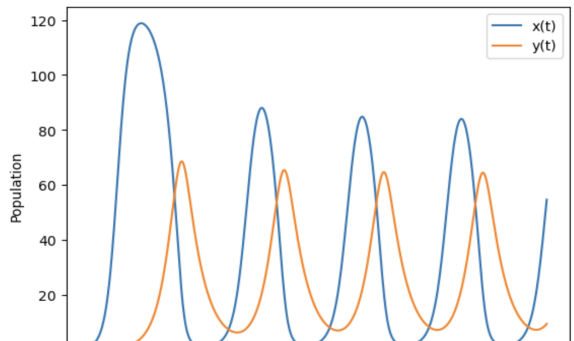
3. Pick points near the equilibrium points solved in 2). (Or you can use initial conditions (x=25, y = 20), (x=1, y=1) , (x=20, y=2).

a) Plot a simulation of the two populations as a function of time.

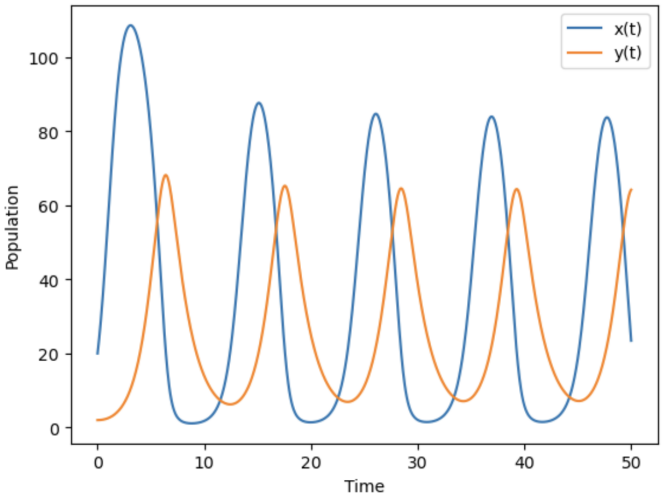
(x=25, y = 20)



(x=1, y=1)

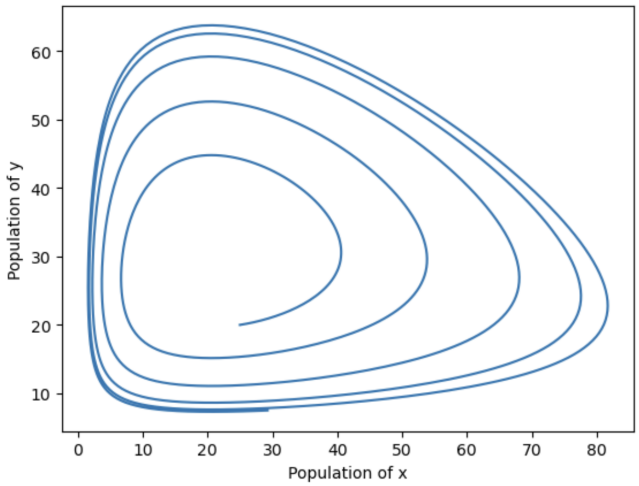


(x=20, y = 2)

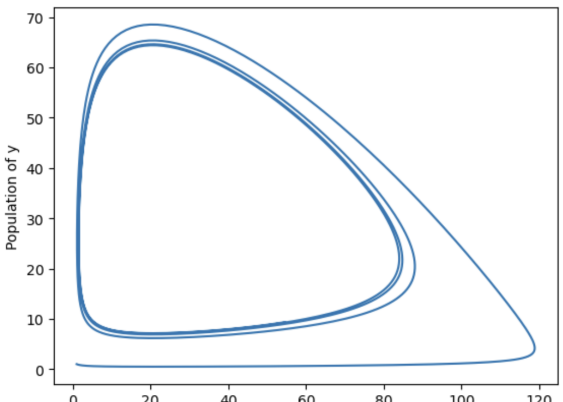


b) Plot a simulation of the two populations plotted against each other.

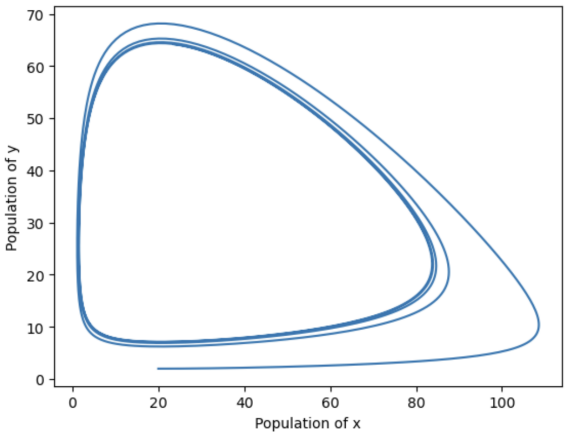
(x=25, y = 20)



(x=1, y = 1)



(x=20, y = 2)


70
60
50
40
30
20
10
0
Population of y
0
20
40
60
80
100
Population of x