Si100 Project Report - Visualization and analysis of epidemic in Shanghai, 2022

Team member: 张敏慎，王婧婷，乔宇堂

大纲

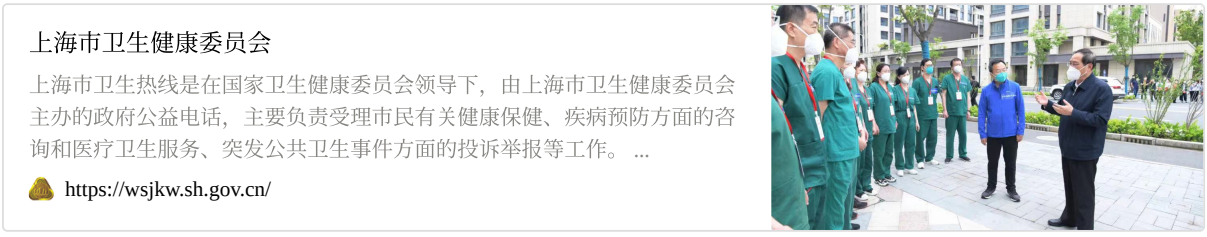
Data Gathering

We use the method of Web crawlers to gather the data we need to use from the websites below. Considering the **reliability** of it, we found the government websites.

中华人民共和国国家卫生健康委员会

中华人民共和国国家卫生健康委员会官方网站

http://www.nhc.gov.cn/



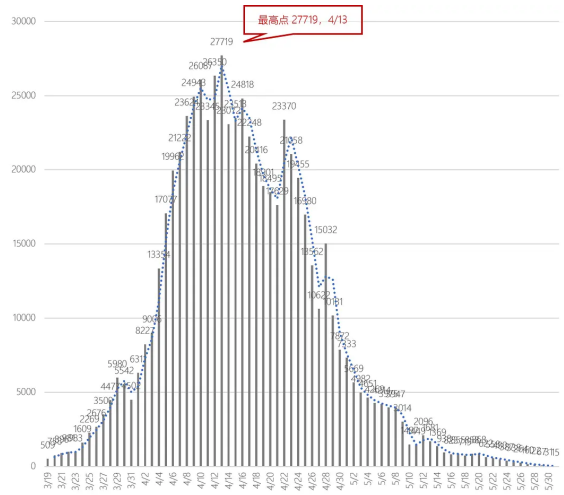
Data Cleaning

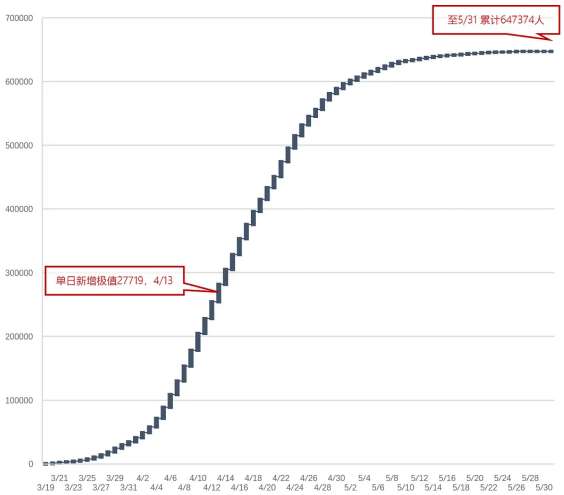
We clean the data by the following code

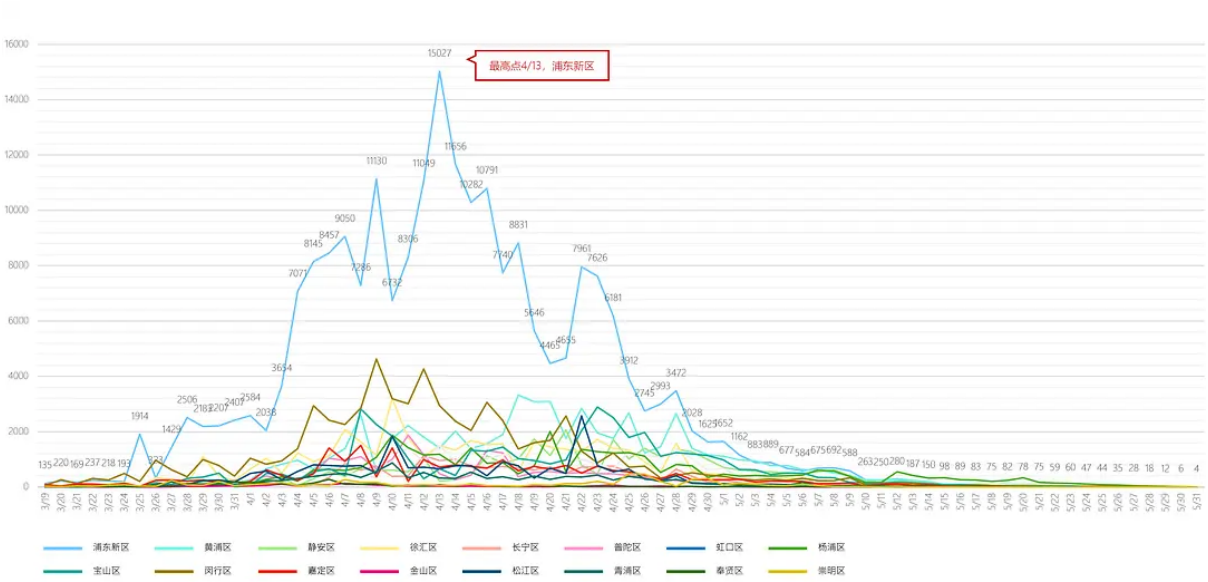
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('各区情况.csv')
# clean step 1: define the fuction to clean NaN

def nan(x):
    if np.isnan(x):
        return 0
    else:
        return x


# apply the map
df = df.applymap(nan)
# clean step 2: fill the NaN in '管控新增阳性.csv' and '风险新增阳性.csv'
guankong = pd.read_csv('管控新增阳性.csv')
fengxian = pd.read_csv('风险新增阳性.csv')
# use interpolate() function
#we can easily fill the empty data very close to real situation
guankong = guankong.interpolate()
fengxian = fengxian.interpolate()
df.to_csv('各区情况(cleaned).csv', encoding="utf-8")
guankong.to_csv('管控新增阳性(cleaned).csv', encoding="utf-8")
fengxian.to_csv('风险新增阳性(cleaned).csv', encoding="utf-8")
```

Data Analysis

We first get a brief impression about the epidemic:





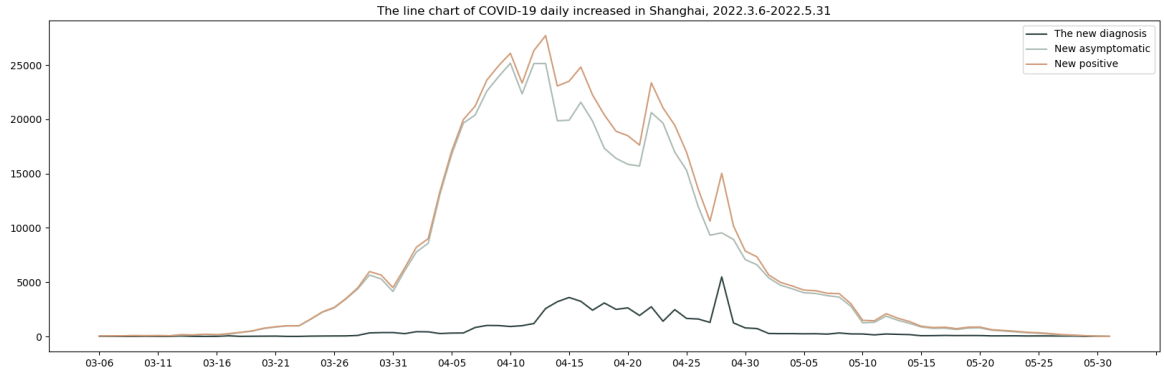


Daily trend

Line Chart

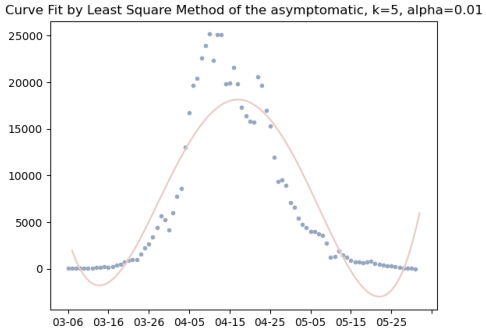
The line chart of COVID-19 situation in Shanghai from 2022.3.6 to 2022.5.31.

The three lines are respectively the new diagnosis, new asymptomatic and new positive.

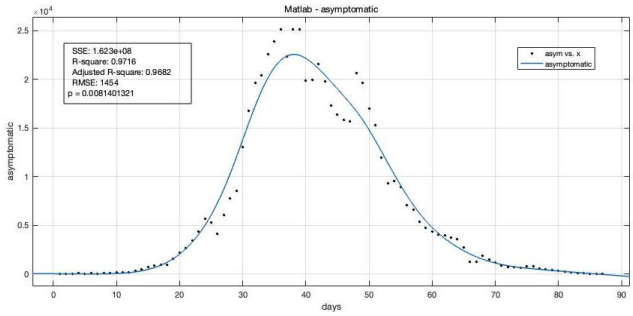


Curve fit by least Square Method

$$f(x) = 4.21 * 10^3 - 1.71 * 10^3 x + 136x^2 - 2.73x^3 + 1.60 * 10^{-2} x^4 3.53 * 10^{-9} x^5$$



Matlab Curve Fitting Smoothing Spline



We can see that by least Square Method, the result is quite **bad and vague**, so we change the model to a more specific one —— SEIRD model.

SEIRD Model

Variable Declaration and Basic assumptions

Let S be the number of susceptible people, E be the number of people already exposed to the virus but having no symptoms, I be the number of people infected and are showing symptoms, R be the number of people recovered, and D be the number of death.

To control the overall number of people be non-changeable, we assume:

$$S + E + I + R + D = Constant \coloneqq P$$

Also we introduce some assistance variable:

Let C be the power of virus spreading, indicating how much people will be infected in one day:

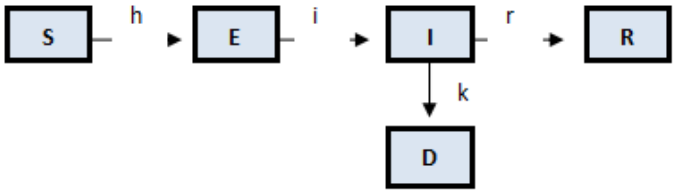
$$C = a \cdot b \cdot P$$

Let R be the propagation rate, indicating the number of people get infected by one infected people in expectation.

$$R = abP \cdot \frac{1}{r} = C \cdot \frac{1}{r}$$

Infection Chain

We have the diagrammatic graph indicating the way virus is spreading



While h indicates the speed or some kind of "rate" of the spreading of the virus, this represents how fast or how easy the spread is. To relate this to more variable, we let:

$$h = a \cdot b \cdot I$$

While a means the contact rate, it numerically equals to the possibility of closely contacting with a infected person.

b **means the efficiency of infection**, it numerically equals to the possibility of successfully infect a susceptible people after contact with him closely, this variable is related to virulence of the virus, whether people are wearing masks, etc. We can vary b to simulate how well people are protecting themselves(wear nothing, one mask, two masks…).

Back to the chain, i **represents the possibility of a exposed people turning into an infected people.**

r represents the possibility of an infected people turning into a recovered people.

k represents the possibility of an infected people die in one day(the possibility will accumulate when day passes by).

Differential equations

$$\begin{cases} \frac{dE}{dt} = h \cdot (P - E - I - R - D) - i \cdot E \\ \frac{dI}{dt} = i \cdot E - (r + k) \cdot I \\ \frac{dR}{dt} = r \cdot I \\ \frac{dD}{dt} = k \cdot I \end{cases} \tag{1}$$

Assume all the variable change discretely by the unit of $\Delta t = 1\ day$, we got the following equation:

$$\begin{cases} E_{t+1} = E_t + h \cdot (P - E_t - I_t - R_t - D_t) - i \cdot E_t \\ I_{t+1} = I_t + i \cdot E_t - (r + k) \cdot I_t \\ R_{t+1} = R_t + r \cdot I_t \\ D_{t+1} = D_t + k \cdot I_t \end{cases} \tag{2}$$

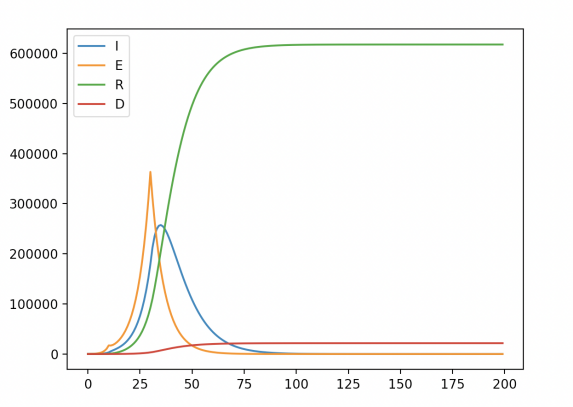
Having this equation, we can calculate all the statistics by knowing the following variables:

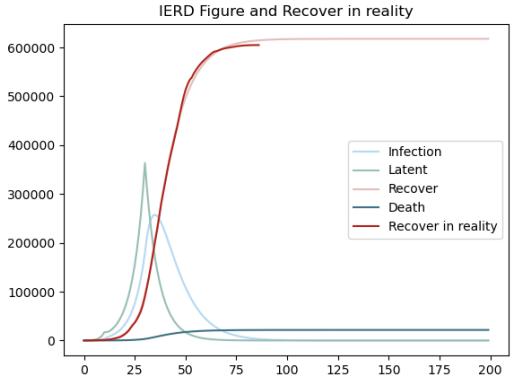
$$E_1, I_1, R_1, D_1, P, a, b, r, k, i$$

So we have crucial part of python code:

```
...

P = 24894300    # 总人数
k = 0.0035  # 日死亡率
i = 1 / 7   # 日转阳率
r = 1 / 10  # 日恢复率
a = 1 / P   # 接触率
b = 0.6     # 感染效率
C = a * b * P   # 传染能量C
R = C / r   # 传播系数

...

# 差分方程
for t in range(T - 1):
    # 前10天，大家没有防备
    if t < 10:
        a = 10 / P
        b = 0.4
    # 10-30天，大家戴上了口罩
    elif 10 <= t < 30:
        a = 3.4 / P
        b = 0.2
    # 隔离在家
    else:
        a = 0.02 / P
        b = 0.01
    h = a * b * I[t]
    E[t + 1] = E[t] + h * S[t] - i * E[t]
    I[t + 1] = I[t] + i * E[t] - (r + k) * I[t]
    R[t + 1] = R[t] + r * I[t]
    D[t + 1] = D[t] + k * I[t]
    S[t + 1] = P - I[t + 1] - E[t + 1] - R[t + 1] - D[t + 1]
    print("第%d天:" % (t + 1), end="")
    print("新增感染者:%d" % (i * E[t]), end="\n")

# 绘制图像
# 预测值
...
```

We get the following result:





So we got the biological parameter of COVID-19 spreading in Shanghai, and the conclusion in detail will be shown in the **Conclusion** part.

Data Interpretation

Critical Date

- On March 28th, Pudong sealed itself(封城)
- On April 11, Puxi sealed itself

With a perspective of the whole city

- Shanghai has 647,374 positive cases overall
- An average of 8,748 people are infected every day
- The peak of number of increasing infection is 27,719
- Pudong New Area increased 15,027 infected people on April 13, which is the largest number over all districts and all time during Shanghai's Epedemic
- In terms of the number of infected people, Pudong New Area is the worst one, with 227,104 positive cases, but we do need to keep in mind that Pudong New Area has the largest population within all districts
- In terms of proportion of infection, Huangpu District is the worst, with an infection rate of 9.26%
- The city-wide average was 2.60%, which means we can find one infected individual in a group of roughly 40 people

With a perspective of every district

- In the 74 days, **Pudong New Area had the largest number of new positive people on the day for most times(26)**. The second was Yangpu District, a total of 19 times.
- From 3/27 to 5/2, Pudong New Area ranked first in the number of new people in Shanghai for 37 consecutive days. It can be said that throughout April, the epidemic in Shanghai was pushed up by

Pudong, the worst-hit area in the city.

- At the beginning of the epidemic, Pudong and Minhang took the lead alternately. After that, Pudong firmly occupied the first place. From May 1st, Huangpu began to be serious. Since the middle of May, Yangpu has become the leader.
- April 13 was the highest number of new cases in the single district of this epidemic, with 15,027 cases in Pudong New Area.

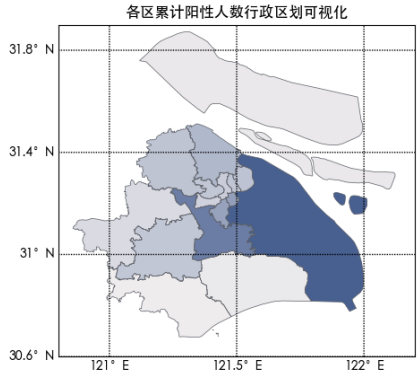
Visualization

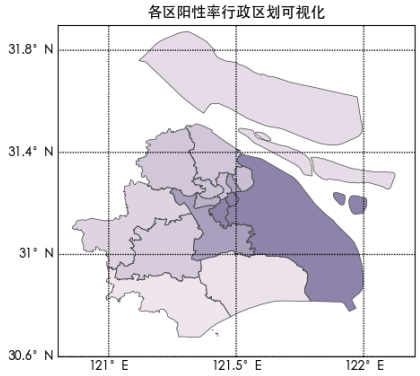
- Visualization is an important part in data analysis and processing, it can help us have a more direct recognization on the connection of every data you have gathered.
- With the help of the library ‘Matplotlib’, we tried many different chart forms to best **show the relationship of our data**, and provide support for our final conclusion or assumptions. Including **histogram**, **column diagram**, **map**, **line chart** and **scatter**.
- Besides, we also consider the help of color. To show the result more visually, we designed a function to **customize our sequential color**. To ensure it to work well, we consider the number distribution and design a function to do more than normal **normalization**.

```
def ColorBar(color0, color1, points: list):
    # 先从16进制转成十进制
    color0 = [int(color0[1:3], 16), int(color0[3:5], 16), int(color0[5:7], 16)]
    color1 = [int(color1[1:3], 16), int(color1[3:5], 16), int(color1[5:7], 16)]

    # 归一化
    deno = max(points)
    mean = np.mean(points)
    normpoints = []
    for num in points:
        if num > mean:
            num = 0.5 + 2 * (num - mean) / deno
        else:
            num = 0.5 * num / mean
        if num > 1:
            normpoints.append(1)
        else:
            normpoints.append(num)

    color = []
    for num in normpoints:
        temcolor = []
        for i in range(3):
            temcolor.append(int(color0[i] + num * (color1[i] - color0[i])))  # 平均计算
        str = '#' + hex(temcolor[0])[2:] + hex(temcolor[1])[2:] + hex(temcolor[2])[2:]   # 转成16进制
        color.append(str)
    return color
```

Visualization about infection rate of every district

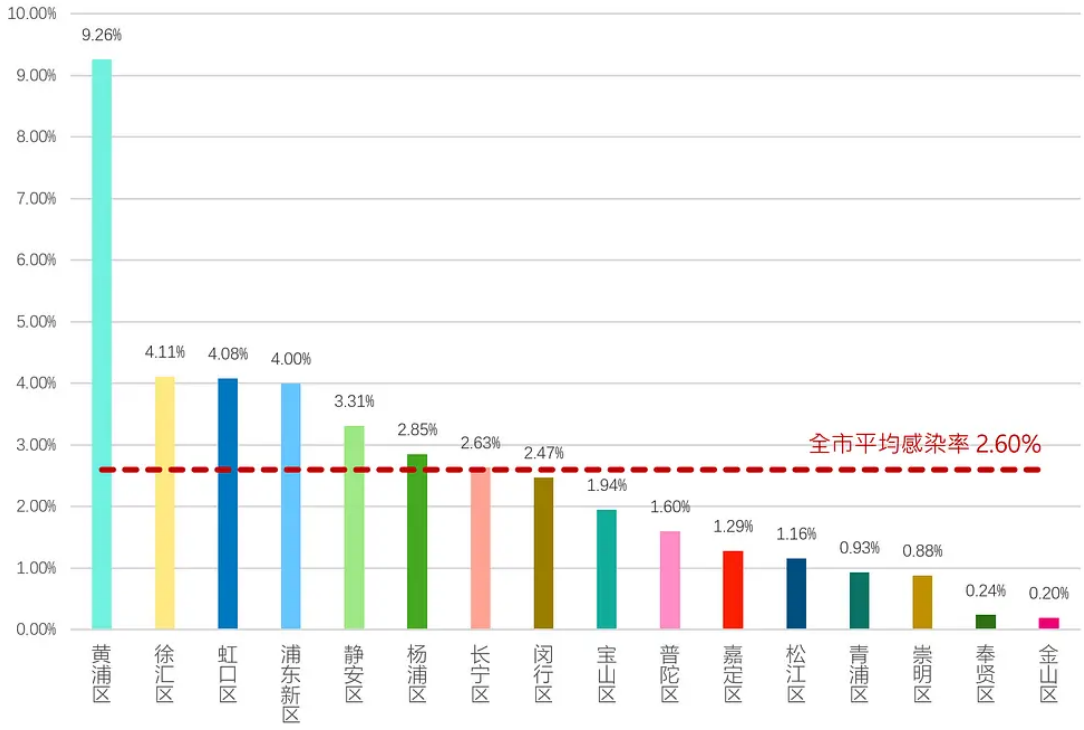
Firstly, we can have a basic impression of the whole situation in every district with the map.



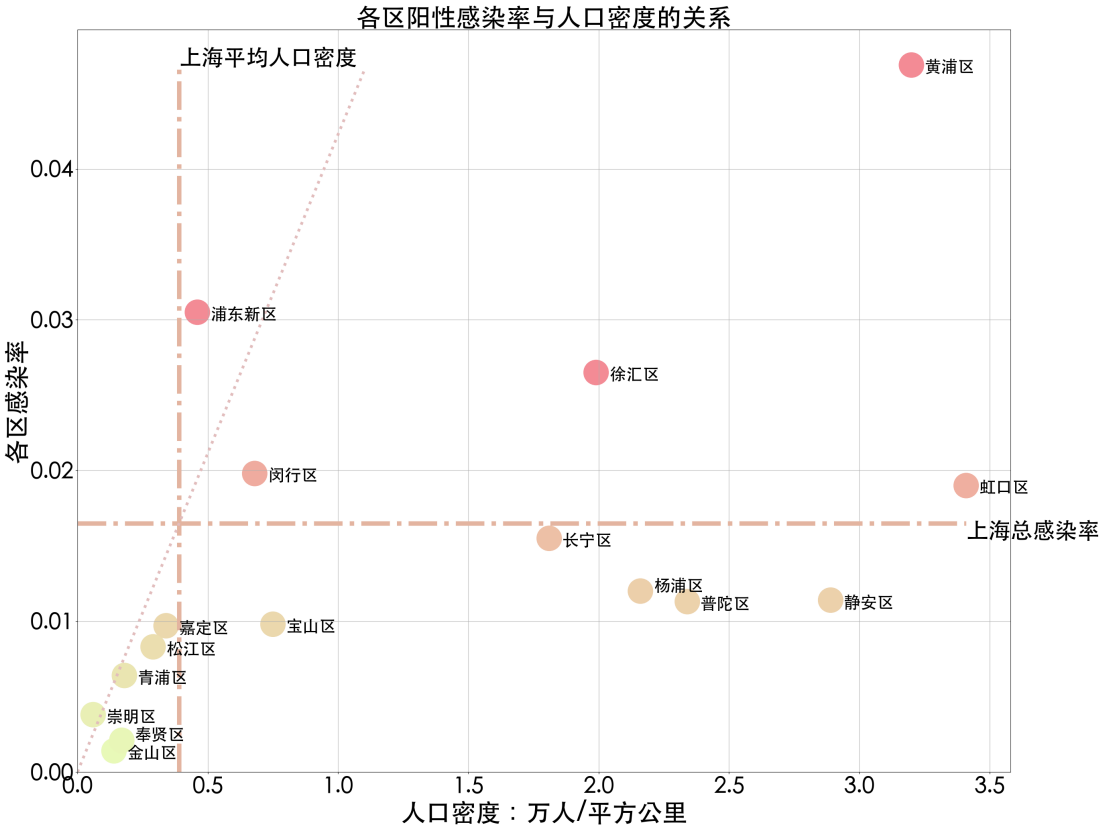


Summary about this visualization:

- Huangpu District had the highest infection rate among all districts of the city, at roughly 9.26%, or we can say, nearly one in 10. This means that in Huangpu district, almost every building, every class, every extended family, someone will be infected.
- Although the total number of positive infections in Pudong New Area is the highest in the city, the infection rate is not very prominent due to the large number of adults in the area.
- The average infection rate in Shanghai was 2.60 percent. Six districts were above average and 10 were below average, indicating that the epidemic was not evenly distributed.



Infection rate and population density visualization



We plot the relationship between infection rate and population density, here's some key information we need to know before getting full information from the graph.

Key Information:

- The first quadrant represents a large population with high infection rates
- The second quadrant represents low population density but high infection rates
- The third quadrant represents low population density and low infection rates
- The fourth quadrant represents high population density and low infection rates
- The Line we draw on this graph indicates the average situation of the whole city, while the district above this line indicates that the district has higher infection/population ratio and vice versa.

Summary about this visualization:

- Although the infection rate of Huangpu district is high, its population density is also high, and the overall ratio is lower than the average situation of Shanghai
- Pudong New Area is actually the hardest-hit area in Shanghai

- Jing'an district and Putuo District performed quite well during this epidemic, while the infection rate remained low in the case of high population density

Conclusion

To summary, we select crucial conclusion from every single part above:

- By simply curve fitting, we can’t get good result and the result is unreasonable, while the SEIRD model performed way better and way reasonably.
- By the SEIRD model, we can know that the Daily Turn Positive Rate is about 1/7, the Daily Recover Rate is about 1/10, and wearing masks can greatly reduce the Infection factor by up to 10 and 100 times.
- Pudong New Area increased 15,027 infected people on April 13, which is the largest number over all districts and all time during Shanghai’s Epidemic.
- The city-wide average was 2.60%, which means we can find one infected individual in a group of roughly 40 people.
- In the 74 days, **Pudong New Area had the largest number of new positive people on the day for most times(26)**. The second was Yangpu District, a total of 19 times.
- From 3/27 to 5/2, Pudong New Area ranked first in the number of new people in Shanghai for 37 consecutive days. It can be said that throughout April, the epidemic in Shanghai was pushed up by Pudong, the worst-hit area in the city.
- At the beginning of the epidemic, Pudong and Minhang took the lead alternately. After that, Pudong firmly occupied the first place. From May 1st, Huangpu began to be serious. Since the middle of May, Yangpu has become the leader.
- April 13 was the highest number of new cases in the single district of this epidemic, with 15,027 cases in Pudong New Area.
- Huangpu District had the highest infection rate among all districts of the city, at roughly 9.26%, or we can say, nearly one in 10. This means that in Huangpu district, almost every building, every class, every extended family, someone will be infected.
- Although the total number of positive infections in Pudong New Area is the highest in the city, the infection rate is not very prominent due to the large number of adults in the area.
- The average infection rate in Shanghai was 2.60 percent. Six districts were above average and 10 were below average, indicating that the epidemic was not evenly distributed.
- Although the infection rate of Huangpu district is high, its population density is also high, and the overall ratio is lower than the average situation of Shanghai
- Pudong New Area is actually the hardest-hit area in Shanghai
- Jing'an district and Putuo District performed quite well during this epidemic, while the infection rate remained low in the case of high population density