1. **Problem definition and description**

The main target of the project is to i) calculate the mean and standard deviation of the data and ii) to find out the representative probabilistic distribution so that we could compare each graph, and figure out the tendency of the amount of CO2 emission by time.

1. **Core code**

# exclude outliers based on the 6 sigma rule

std\_temp = df[column\_name].std()

mean\_temp = df[column\_name].mean()

lower\_limit = df[column\_name] > mean\_temp - 3 \* std\_temp

upper\_limit = df[column\_name] < mean\_temp + 3 \* std\_temp

df = df[lower\_limit & upper\_limit]

# calculate statistics of the new dataframe

min\_val = df[column\_name].min()

max\_val = df[column\_name].max()

std\_val = df[column\_name].std()

mean\_val = df[column\_name].mean()

# Draw histogram

x\_list = []

step = (max\_val - min\_val) / x\_steps

for i in range(0, x\_steps + 5):

x\_list.append(min\_val + (i - 5) \* step)

plt.hist(df[column\_name], x\_list, density = True, label = 'Histogram')

# Probability Density Function 1: normal distribution graph

x = np.linspace(min\_val - 5 \* step, max\_val + 5 \* step, 1000)

y = (1 / (np.sqrt(2 \* np.pi) \* std\_val)) \* np.exp(-(x-mean\_val)\*\*2 / (2 \* std\_val\*\*2))

plt.plot(x, y, alpha=0.7, label=norm\_label, linewidth = 4)

# Probability Density Function 2: Maximum Likelihood Estimation

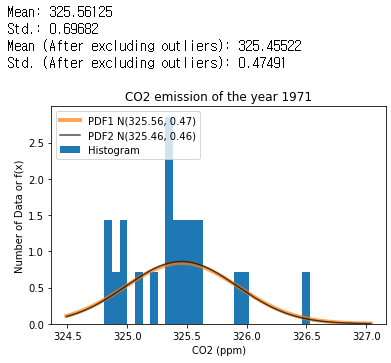
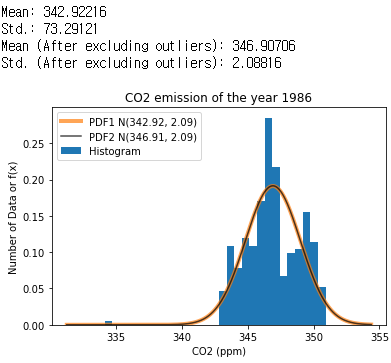
x\_fit = df[column\_name]

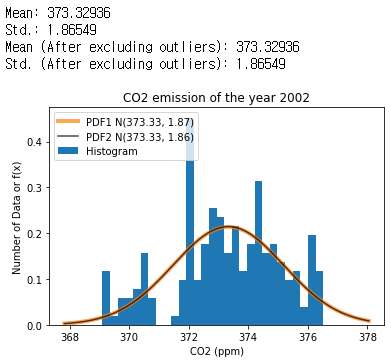
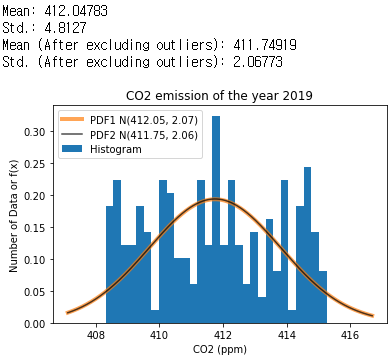
loc, scale = scipy.stats.norm.fit(x\_fit)

y\_fit = (1 / (np.sqrt(2 \* np.pi) \* scale)) \* np.exp(-(x-loc)\*\*2 / (2 \* scale\*\*2))

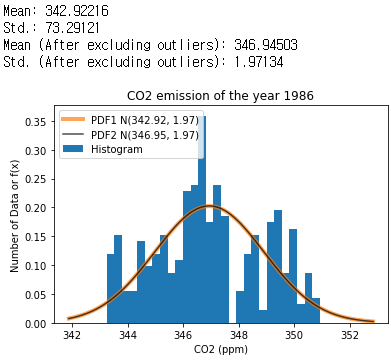
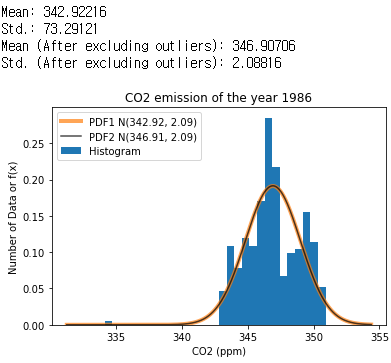
plt.plot(x, y\_fit, alpha=0.7, label=fit\_label, color = 'black')

1. **Results and plots**

**Figure 1, 2, 3, 4(Clockwise from the upper left). CO2 emissions of the year 1971, 1986, 2019, 2002**



**Figure 5. Excluded remained outlier after applying 6-sigma rule, manually**

1. **Discussion**

By observing the plotted data into those graphs, we can strongly argue the analysis below:

1. The probability density function figured out from the maximum likelihood estimation (MLE) will finally converge to the normal distribution that follows the mean and standard deviation of the data, as we can see in all figures 1, 2, 3, 4.
2. Sometimes the 6-sigma rule to exclude outliers would leave some values at the edge, which would have not affected the distribution of the graph, as we can see in figure 5. However, excluding the remains manually did not change the distribution very significantly.
3. **Refernces**

The SciPy Community, 2008-2021. *scipy.stats.rv\_contin  
uous.fit*, Available online: https://docs.scipy.org/doc/sci  
py/reference/generated/scipy.stats.rv\_continuous.fit.html (accessed on 17 March 2021).