Project 4 Experimentation

Name: Peng Yan

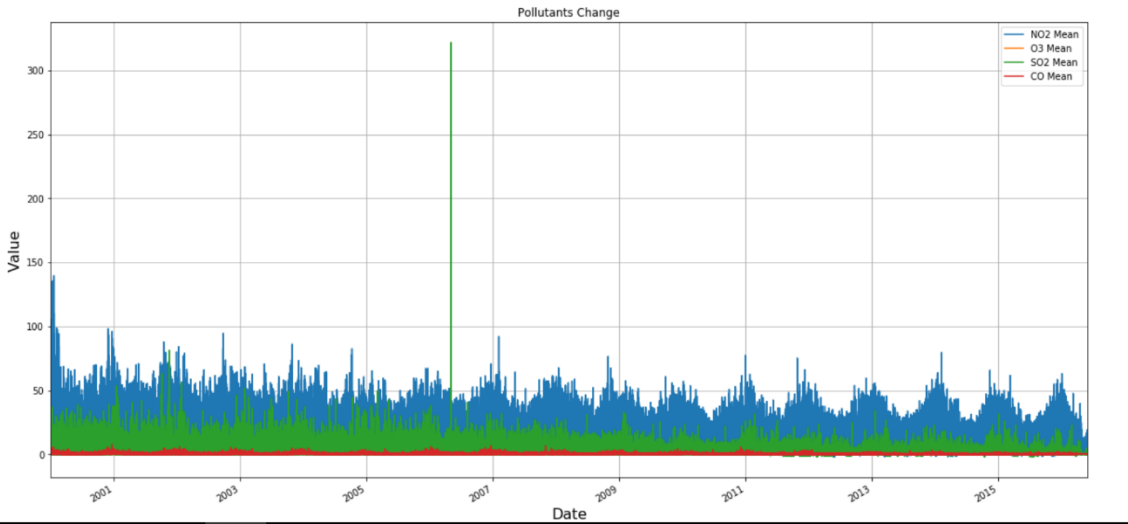
Part 1: Designing an Experiment

Write an experiment to test a specific question about an aspect of the effectiveness of visualizations. Note that this experiment should address a specific question that you can answer by measuring people's performance in a certain task.

I use the data set come from https://www.kaggle.com/sogun3/uspollution/data. This dataset deals with pollution in the U.S. There is a total of 28 fields. The four pollutants (NO2, O3, SO2 and O3) each has 5 specific columns. I try to use as much as possible data to build a graph. This data set gathered four major pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 – 2016. Then I look 5 classmates give me a feedback how many information they could get form this visualization. I design the scores from 0 to10. 0 means nothing could get from the visualization and 10 means could get enough information.

I generate the visualization as below. I look for some classmates help me test this graph and tell me how many information they could get from this visualization.

```
In [13]: ax=dfH.plot(x = "Date Local", y = ["NO2 Mean", "O3 Mean", "SO2 Mean", "CO Mean"],figsize=(20,10), grid=True,title="Pollu
         ax.set_xlabel("Date", fontsize=16)
         ax.set_ylabel("Value", fontsize=16)
         plt.show()
```



- Research Question (What question will you address?)

  I try to build a complicated visualization that would packed with as much data as possible. My research question will try to talk whether a as much data as possible visualization could provide more information to users or much better than a simple visualization.

- Hypotheses (What do you think will happen and why?)

  I think if there are as much data or information as possible in the visualization that could improve the effectiveness of the visualizations and the graph could provide more information to users. Because if there are big enough data in the visualization, user could get more useful information form this kind of visualization.

- Independent Variables (What will you change? What are the levels of each?)

  In this visualization, the independent variable is the time and the level is every day from 2000 to 2016

- Dependent Variables/Measures (What will you measure?)

  Dependent variables are the mean value of four pollutants (NO2, O3, SO2 and O3) at each day.

- Control Variables (What else do you need to account for? How will you do it?)

  I want to know these four major pollutants change trend from 2000 to 2016

- Description of the Stimuli (What will the participant see? Can be an annotated sketch or a verbal description. Include the source of any data needed)

  I will show the graph to the participant. At the same time, I will tell them what the graph shows, the independent variable and Dependent variables.

- Experimental Procedure (What will the participant do? Please describe this using a step-by-step procedure and include any details necessary to conduct the experiment)

  First, I will show the graph to the participant.

  Second, I will tell them what the graph shows, the independent variable and Dependent variables.

  Third, I want to the participant tell me how many information they get from this visualization.

- Planned Analysis (How will you analyze your dependent variables and why?)

  I try to look for some outliers and find the change trend from 2000 to 2016. I also could the change trend for each year about these four major pollutants.

Part 2: Building the Apparatus

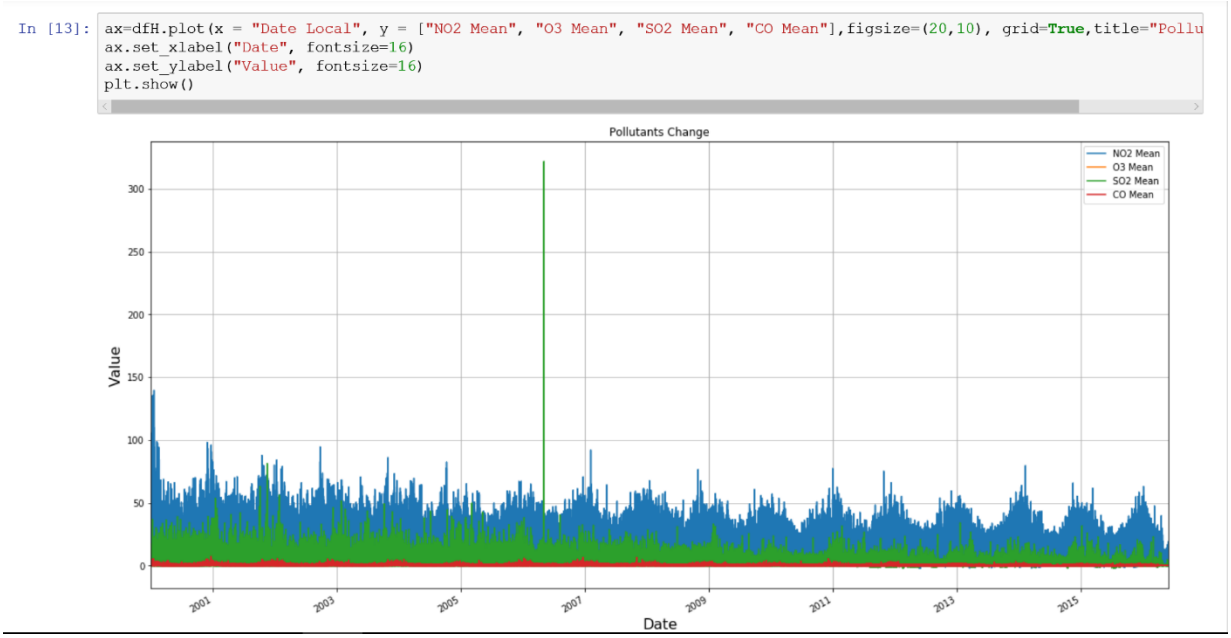
Construct the apparatus you need to conduct this experiment. This can be a program (e.g., a web site or python program) or the full set of questions and stimuli you'd use to test your hypotheses. If you choose to provide the stimuli and questionnaires rather than a program, note in your document how those pieces would fit together.

I wrote a python code for build this graph.

```
import pandas as pd
import numpy as np
import sys
import matplotlib.pylab as plt
```

```
%matplotlib inline
import pylab as pl
from pandas import *

dfH= pd.read_csv("usp.csv")  #read the file
dfH
dfH['Date Local']=pd.to_datetime(dfH['Date Local'])  #change the date type
print(dfH.head(10))
#plot the line graph
ax=dfH.plot(x = "Date Local", y = ["NO2 Mean", "O3 Mean", "SO2 Mean", "CO
Mean"],figsize=(20,10), grid=True,title="Pollutants Change")
ax.set_xlabel("Date", fontsize=16) #X axis
ax.set_ylabel("Value", fontsize=16) #Y axis
plt.show()
```



Part 3: Conducting the Study

Use your apparatus to collect data from at least 5 different people. In your document, provide a visualization of the resulting data that helps answer your research question and briefly describe what you find.

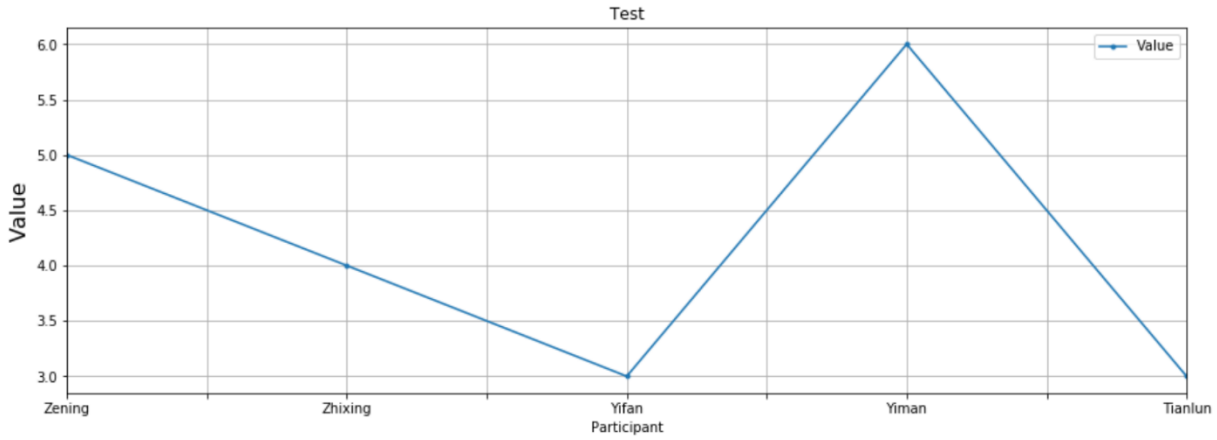
I design the scores from 0 to10. 0 means nothing could get from the visualization and 10 means could get enough information.

Below is my code for visualization of the resulting data.

```
import pandas as pd
import numpy as np
import sys
import matplotlib.pylab as plt
```

```
%matplotlib inline
import pylab as pl
from pandas import *

#build the data
data = {'Participant': ['Zening', 'Zhixing', 'Yifan', 'Yiman', 'Tianlun'],
        'Value':   [5, 4, 3, 6, 3]
        }
df = pd.DataFrame(data)
print(df)
#plot the data
ax=df.plot(x = 'Participant', y = 'Value',figsize=(15,5), grid=True,title="Test",style='.-')
ax.set_xlabel("Participant", fontsize=16)
ax.set_ylabel("Value", fontsize=16)
```



Part 4: Inferential Analysis

Use either inferential or bayesian methods to analyze the outcomes of your experiment in accordance with your planned analysis in part 1. In your document, describe what you found and what it tells you about your research question.

Through our result we can see too much data in the visualization sometime cannot improve the effectiveness. Because sometimes it difficult to put whole big data in one graph. If we put a big data into one graph, the graph may be not clear to users. Compare with my hypotheses, I could see we get contrary conclusion.