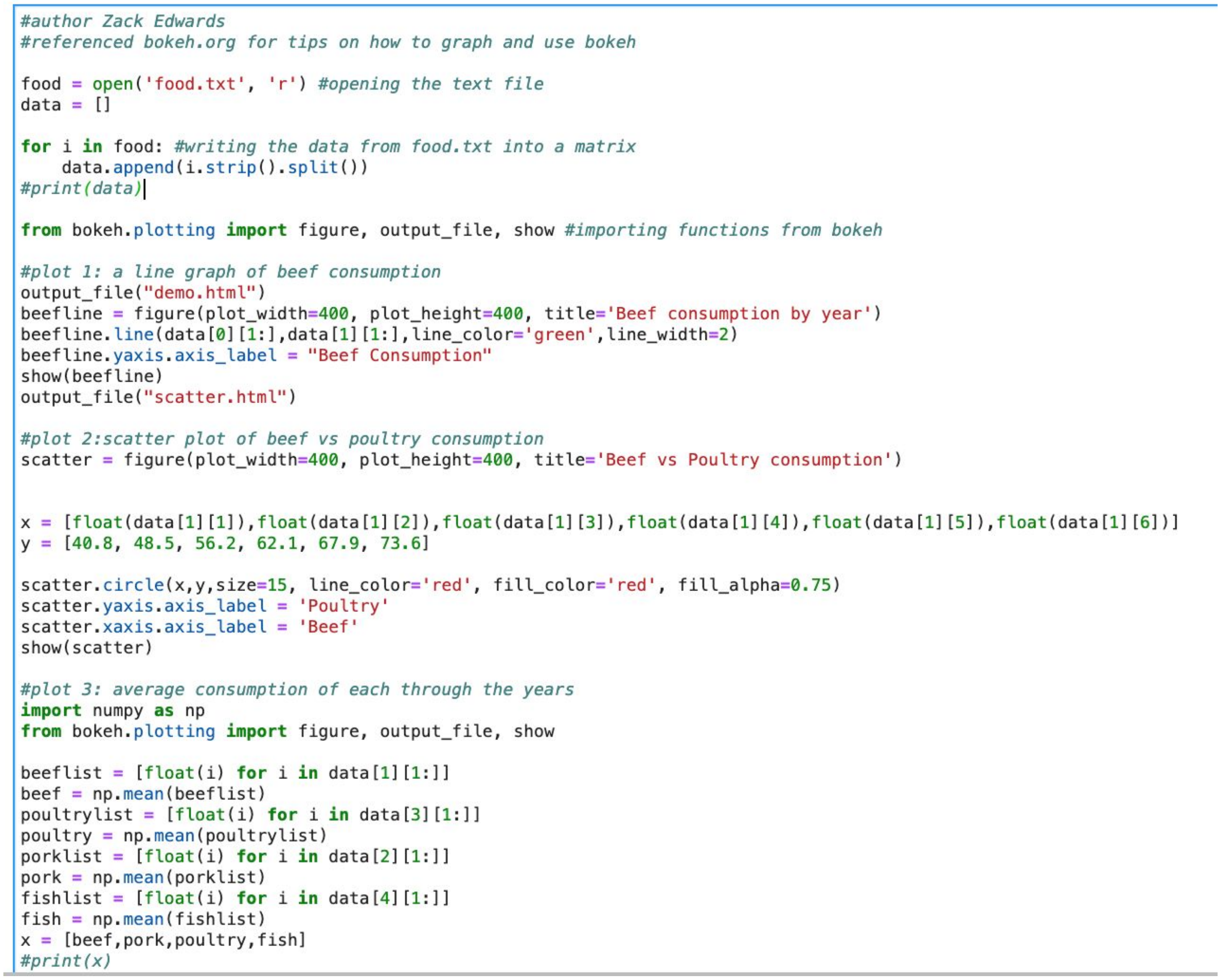
Zack Edwards

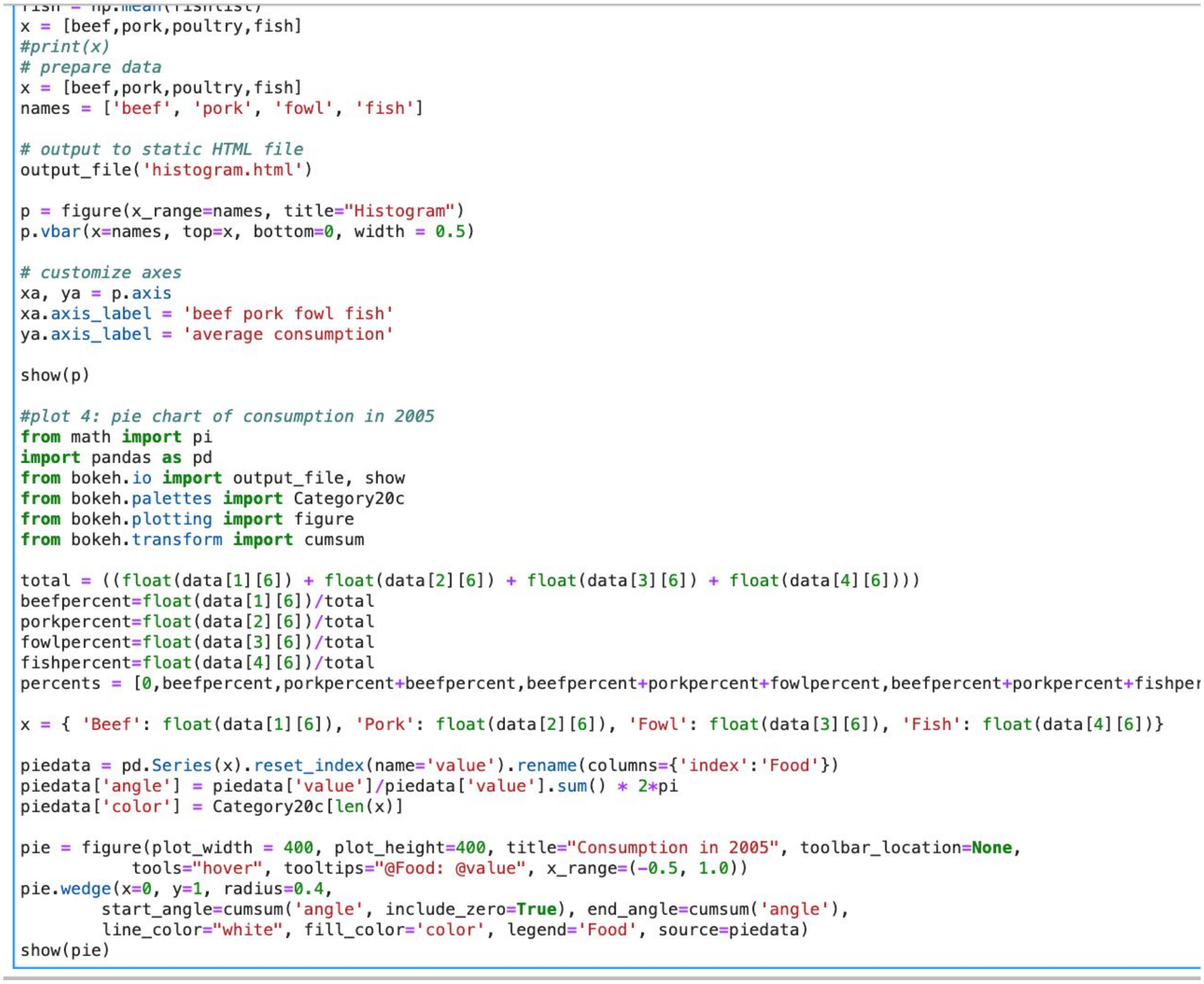
Questions:

1) What is text mining preparation?
    a) Text mining is a process which is used to clean and analyze data. Text mining preparation is the act of cleaning and organizing unstructured data into individual key words or whatever form you wish the data to take which can then be used for text mining analysis.
2) What are the differences between Wordnet and a dictionary?
    a) A dictionary contains all the words in a language and their definitions. The Wordnet lexicon contains all words in 200 languages and their limitless semantic associations with other words, such as synonyms, antonyms and more. In essence it is a library of hidden connections and meanings within words. It can be used to analyze words while doing text mining.
3) What is “part of speech tagging”?
    a) Part of speech tagging is another tool to be used during text mining. It is used to analyze words in a block of text by tagging each word with a code indicating its part of speech (i.e noun verb adjective adverb). It does this based on both context and definition.

Writing the code:



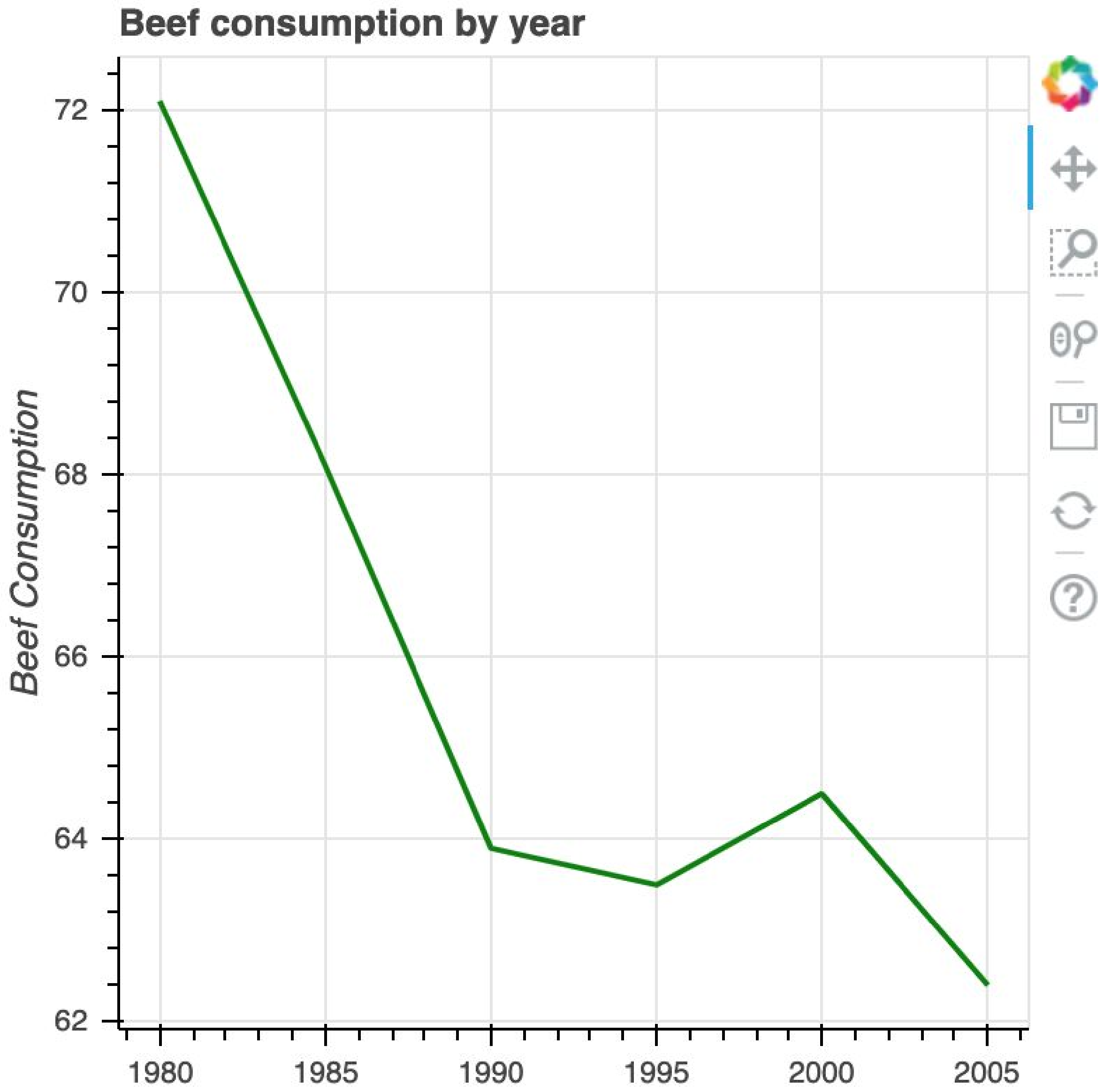
```
#author Zack Edwards
#referenced bokeh.org for tips on how to graph and use bokeh

food = open('food.txt', 'r') #opening the text file
data = []

for i in food: #writing the data from food.txt into a matrix
    data.append(i.strip().split())
#print(data)

from bokeh.plotting import figure, output_file, show #importing functions from bokeh

#plot 1: a line graph of beef consumption
output_file("demo.html")
beefline = figure(plot_width=400, plot_height=400, title='Beef consumption by year')
beefline.line(data[0][1:],data[1][1:],line_color='green',line_width=2)
beefline.yaxis.axis_label = "Beef Consumption"
show(beefline)
output_file("scatter.html")

#plot 2:scatter plot of beef vs poultry consumption
scatter = figure(plot_width=400, plot_height=400, title='Beef vs Poultry consumption')


x = [float(data[1][1]),float(data[1][2]),float(data[1][3]),float(data[1][4]),float(data[1][5]),float(data[1][6])]
y = [40.8, 48.5, 56.2, 62.1, 67.9, 73.6]

scatter.circle(x,y,size=15, line_color='red', fill_color='red', fill_alpha=0.75)
scatter.yaxis.axis_label = 'Poultry'
scatter.xaxis.axis_label = 'Beef'
show(scatter)

#plot 3: average consumption of each through the years
import numpy as np
from bokeh.plotting import figure, output_file, show

beeflist = [float(i) for i in data[1][1:]]
beef = np.mean(beeflist)
poultrylist = [float(i) for i in data[3][1:]]
poultry = np.mean(poultrylist)
porklist = [float(i) for i in data[2][1:]]
pork = np.mean(porklist)
fishlist = [float(i) for i in data[4][1:]]
fish = np.mean(fishlist)
x = [beef,pork,poultry,fish]
#print(x)
```



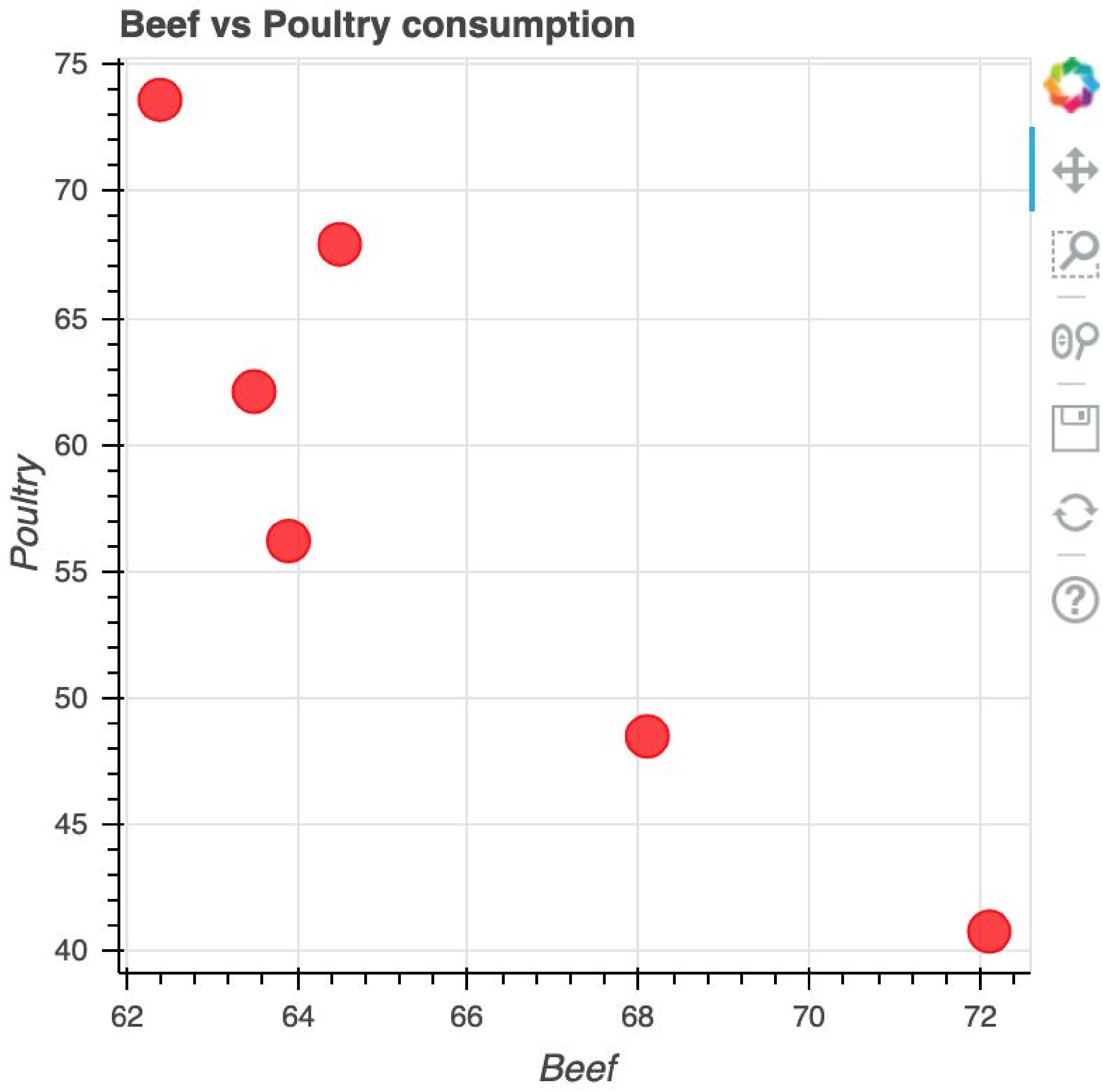
```
fish = np.mean(fishlist)
x = [beef,pork,poultry,fish]
#print(x)
# prepare data
x = [beef,pork,poultry,fish]
names = ['beef', 'pork', 'fowl', 'fish']

# output to static HTML file
output_file('histogram.html')

p = figure(x_range=names, title="Histogram")
p.vbar(x=names, top=x, bottom=0, width = 0.5)

# customize axes
xa, ya = p.axis
xa.axis_label = 'beef pork fowl fish'
ya.axis_label = 'average consumption'

show(p)

#plot 4: pie chart of consumption in 2005
from math import pi
import pandas as pd
from bokeh.io import output_file, show
from bokeh.palettes import Category20c
from bokeh.plotting import figure
from bokeh.transform import cumsum

total = ((float(data[1][6]) + float(data[2][6]) + float(data[3][6]) + float(data[4][6])))
beefpercent=float(data[1][6])/total
porkpercent=float(data[2][6])/total
fowlpercent=float(data[3][6])/total
fishpercent=float(data[4][6])/total
percents = [0,beefpercent,porkpercent+beefpercent,beefpercent+porkpercent+fowlpercent,beefpercent+porkpercent+fishper

x = { 'Beef': float(data[1][6]), 'Pork': float(data[2][6]), 'Fowl': float(data[3][6]), 'Fish': float(data[4][6])}

piedata = pd.Series(x).reset_index(name='value').rename(columns={'index':'Food'})
piedata['angle'] = piedata['value']/piedata['value'].sum() * 2*pi
piedata['color'] = Category20c[len(x)]

pie = figure(plot_width = 400, plot_height=400, title="Consumption in 2005", toolbar_location=None,
           tools="hover", tooltips="@Food: @value", x_range=(-0.5, 1.0))
pie.wedge(x=0, y=1, radius=0.4,
        start_angle=cumsum('angle', include_zero=True), end_angle=cumsum('angle'),
        line_color="white", fill_color='color', legend='Food', source=piedata)
show(pie)
```

Outputs:

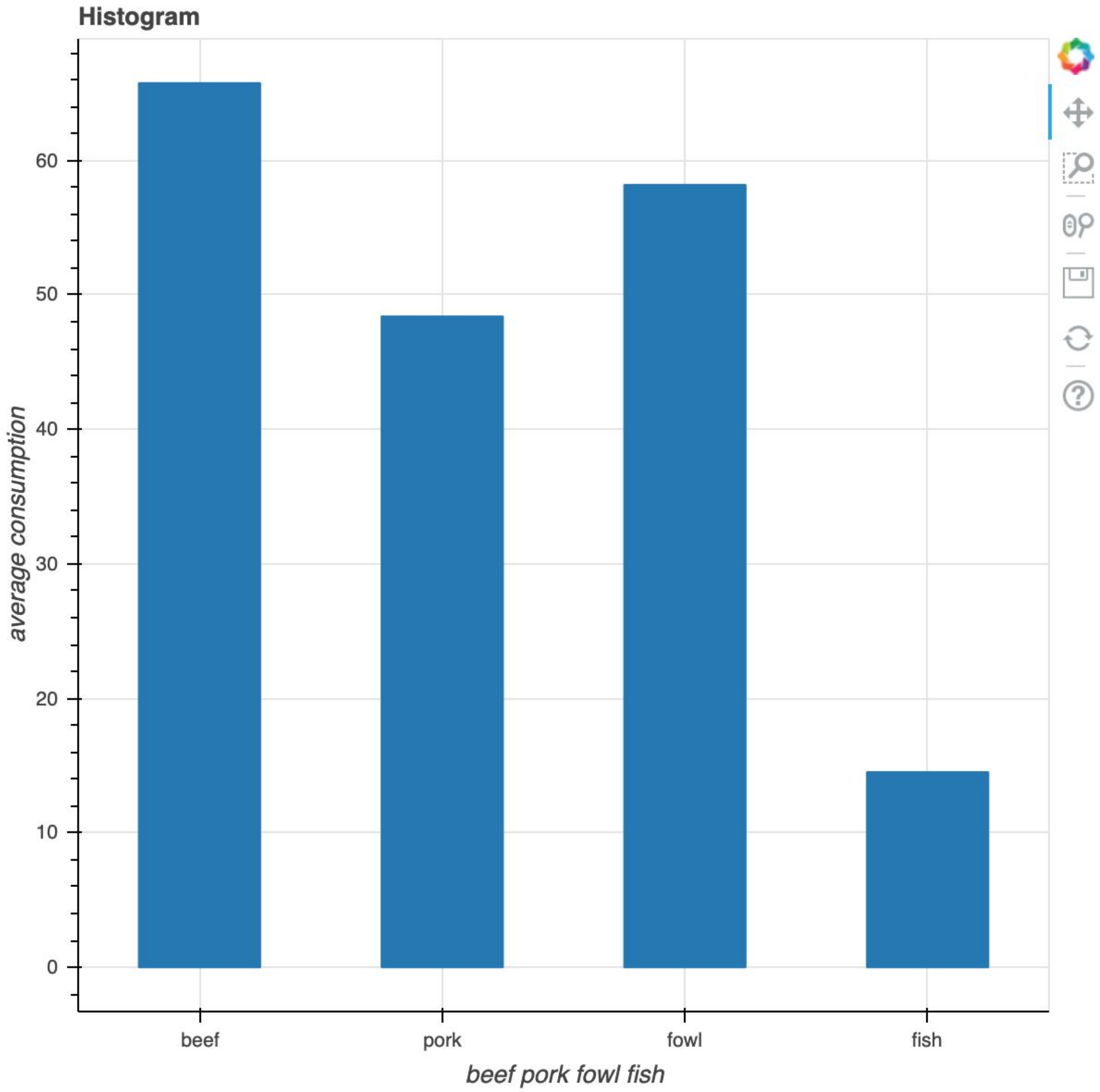
Graph 1


Beef consumption by year
Beef Consumption
72
70
68
66
64
62
1980
1985
1990
1995
2000
2005

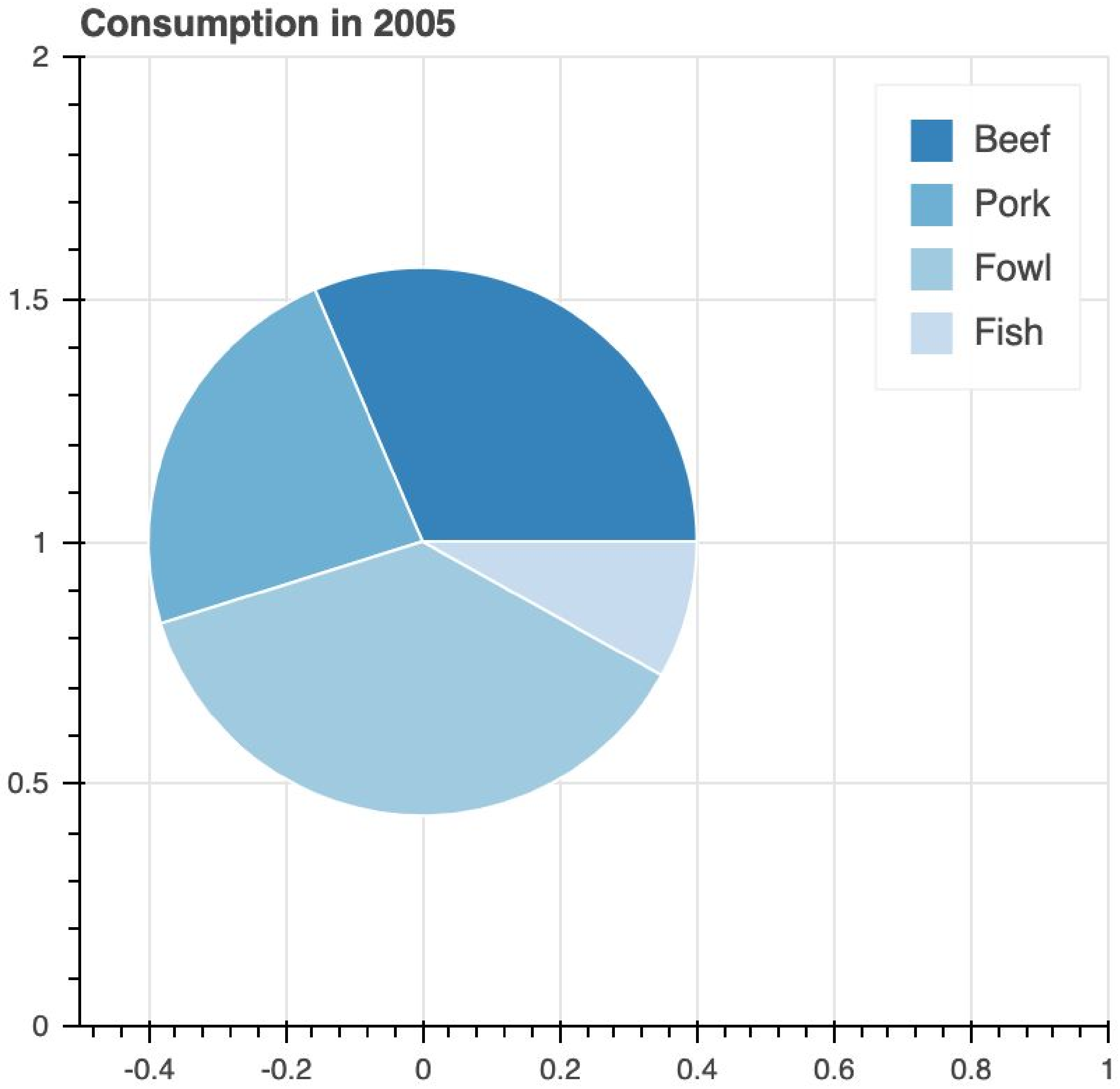
Graph 2


Beef vs Poultry consumption
75
70
65
60
55
50
45
40
Poultry
62
64
66
68
70
72
Beef

Graph 3


Histogram
average consumption
0
10
20
30
40
50
60
beef
pork
fowl
fish
beef pork fowl fish

Graph 4


Consumption in 2005
2
1.5
1
0.5
0
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Beef
Pork
Fowl
Fish