Data Analysis of Pedometer Steps

Fall 2019: Visual Analytics

Sheridan Kamal, Merissa Lissade, Tova Schwartz

Abstract

Exercise is the modern panacea: the cure to obesity, heart disease, and many mental health conditions. Exercise is also the most deprived need; neglected by those who need it most. The cause of this phenomenon is attributed to a host of excuses; for example: high costs, lack of time, and bad weather. To account for these three main excuses we chose to analyze walking, a free form of exercise, and its relation to time and temperature. How elastic is activity level to the daily temperature, how fixed is our daily routine, does our weekly activity level reflect our weekly schedule? We hypothesize that, as an overall trend, days with nice, mild weather will correlate with a higher activity level and days with high, uncomfortable temperatures will correlate with a lower activity level. This will account for the casting of weather as the scapegoat for inactivity. Additionally, we predict that a weekly trend will arise that reflects a daily schedule. If scheduling is truly responsible for the scarcity of physical activity, then the weekly activity will cyclically correlate with a weekly schedule.

Introduction

America is prosperous, GDP is high, and along with this state of abundance comes an obesity epidemic. According to the National Health and Nutrition Examination Survey, one in every three Americans is overweight. According to the Center for Disease Control and

Prevention: one in every four deaths is due to heart disease and 9.4% of Americans suffer from diabetes. The most widely agreed-upon technique for avoiding said statistics is exercise. The cure is simple in theory but provenly difficult in practice.

Exercise is one of the most common and most neglected New Year's resolutions. Gym memberships are costly, schedules are packed, and workouts are inconvenient. Walking, however, is free, is done within the framework of a typical daily schedule, and provides high returns on health benefits. For example; according to Bluearth, walking provides an array of health benefits including, but not limited to, reducing the risk of heart disease and stroke. However, an article published by Wendy Bumgardner cites that the average American walks around 5,000 steps a day. This is way below the number of recommended daily steps of approximately 10,000. If walking is such an idealized exercise why is its neglect so widespread?

Although walking is free, obstructions to this practice do exist. On the East Coast, where this study was conducted, the weather is a common scapegoat for inactivity. Does the weather truly have an impact on the activity level? Is there a correlation between temperature and steps taken? If the weather excuse is valid, then good weather will correlate with higher activity levels and vice versa.

An additional pitfall of walking is the time aspect. Recreational walking is no less time consuming than attending the gym or an exercise class. Only walking as a means of transportation can avoid the pitfalls of 'exercise activities.' If time constraints are a valid excuse, and recreational walking is therefore neglected, then the only walking that is done is as a means of transportation. Following this assumption, an individual's daily number of steps would very accurately reflect a weekly routine. Does walking data reflect a cyclical weekly pattern, correlating to one's weekly schedule? Does walking increase on weekends when one has more free time?

To explore this topic and answer the aforementioned questions we chose to analyze personal data. The data was extracted from the Apple Health App. The data were merged with daily temperature estimates from the NOAA website. The data is a time series lapsing from June 1st, 2019 to September 29th, 2019. The dataset has the following variables: date; steps walked for Merissa, Sheridan, and Tova; and the daily temperature.

Our hypothesis is that the activity level will correlate with temperature and follow a weekly cyclical trend. Since our data is from the summer months, we expect the temperature to be negatively correlated with steps. This is because the intense heat of the summer provides an excuse for East Coasters to not participate in outdoor recreation. Secondly, we expect the steps to follow a cyclical trend mirroring our weekly schedules. During the week we follow a strict schedule that is reflected in our steps, while weekends present greater variability and activity since it is less rigidly timed.

Related Works

The world today is a more health-conscious one than in previous years. Step data has been used for a variety of health studies due to the ease of use and easy accessibility of pedometer step data. There are several related works that also utilize pedometer step data to perform a meta-analysis as we have done in this project.

Richardson et al. 2008 used the pedometer step data gathered from 9 studies (which was taken from 6 databases) to analyze the relationship between activity levels and weight loss without dietary intervention. After analyzing the data of 307 participants (73% of which were female and 27% of which were male) it was determined that the more activity the higher the weight loss as well as the longer the study the higher the weight loss. The weight loss was only a modest amount and would likely be higher with a dietary intervention program as well.

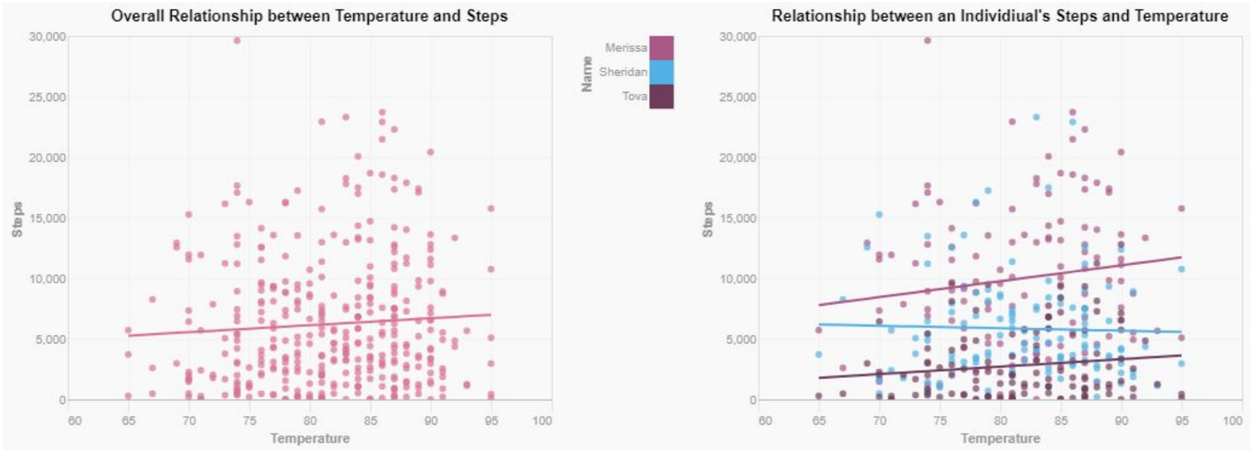
Baskerville et al. 2017 used both pedometers (9 studies) and accelerometers (3 studies) to analyze the relationship between activity levels and HbA1c levels in individuals with Type 2 diabetes. After analyzing the data of 1458 trial participants, it was determined that through the use of either the pedometer or accelerometer overall activity levels increased, but there were no significant differences in HbA1c levels. It could be concluded that although activity levels increased during the use of either the pedometer or accelerometer there is additional intervention needed in order to lower HbA1c levels.

Le Masurier et al. 2005 used the pedometer step data gathered from 6 existing datasets to analyze the activity levels of school-aged children. After separating the data of 1839 participants (1046 of which were female and 793 of which were male) by gender and by grade cluster (elementary grades 1-3, upper elementary grades 4-6, middle school grades 7-9, and high school grades 10-12) and analyzing the data, it was determined that males are more active than females and students in grades 1-6 are more active than students in grades 7-12.

Methodology

Visualization 1, by: Sheridan Kamal

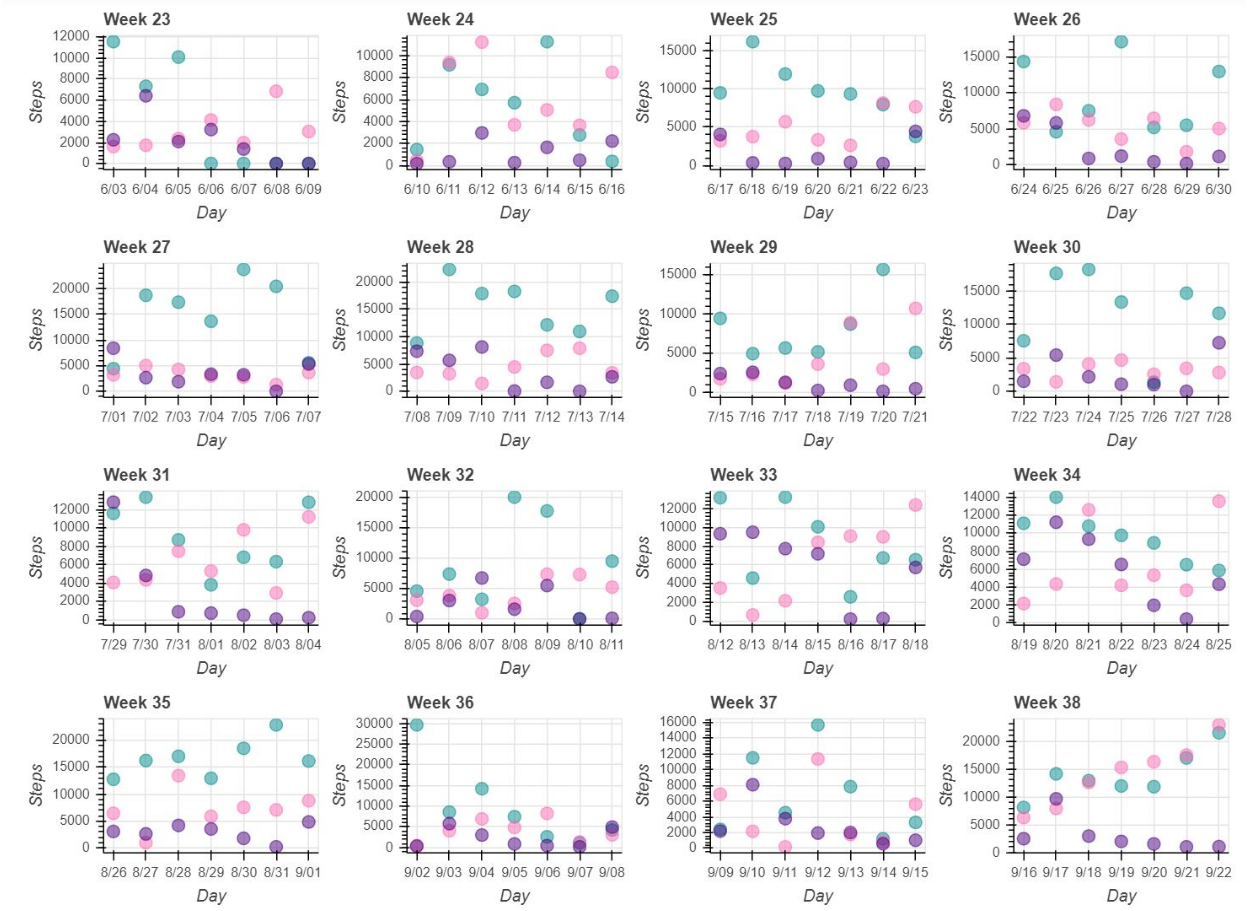
The following visualizations were made using the Altair library in Python3 and seek to explore the effect temperature has on steps, both in general and on an individual basis. Since we are seeking to investigate the relationship between temperature and steps, the visualization that was chosen was a scatter plot layered with a linear trendline because this was the best visualization to display the data. The Python3 code for the project can be found here: https://repl.it/@skamal/CSC83060-FinalProject. The HTML5 code for the interaction can be found here: https://repl.it/@skamal/CSC83060-Final-Project. The fully interactive version of this visualization can be found here: https://csc83060-final-project.skamal.repl.co/. The legend in the middle of the two visualizations can be used to filter the visualizations by name and can be used to compare an individual's relationship between temperature and steps shown on the right to the overall relationship between temperature and steps shown on the left. Hovering over the trendline shows the formula for that trendline and hovering over the data points also tells you the temperature, the amount of steps taken, and the individual's name who that data point represents.



We originally hypothesized that as temperature increases, the amount of steps walked would decrease and would show a negative linear relationship when visualized. However, there seems to be an overall positive linear relationship between temperature and the amount of steps walked as shown by the visual on the left. The visualization on the right shows that when we plot an individual's steps walked against temperature each individual has their own unique relationship between temperature and steps walked. While it is true that Sheridan's step data shows a negative linear relationship with temperature and supports our hypothesis, Merissa's and Tova's step data does not. Merissa's step data and Tova's step data both show a positive linear relationship with temperature although Merissa seems to walk more steps as the temperature increases than Tova does.

Visualization 2, by: Merissa Lissade

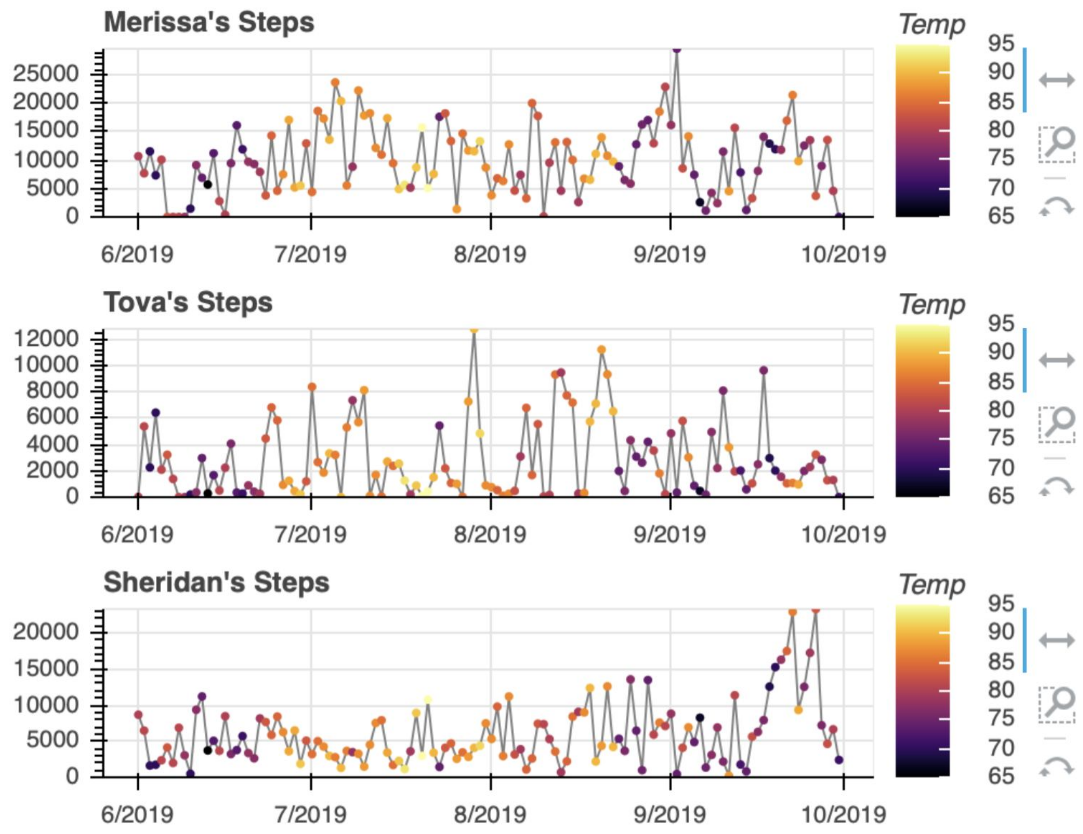
The visualization in this section was created using Python Bokeh Library via Jupyter Notebook. We wondered if there were any cyclic patterns that could be seen on a weekly basis. In this visualization, weeks 23 - 26 include the days in June, weeks 27 - 31 include the days in July, weeks 31-35 include the days in August and weeks 35-38 include the days in September.



We figured given some of us may have certain daily patterns, that maybe this would reflect in our step data as well. However, it can be seen from the visualization that there does not appear to be a particular day of the week or any type of pattern in which any individual walked the most.

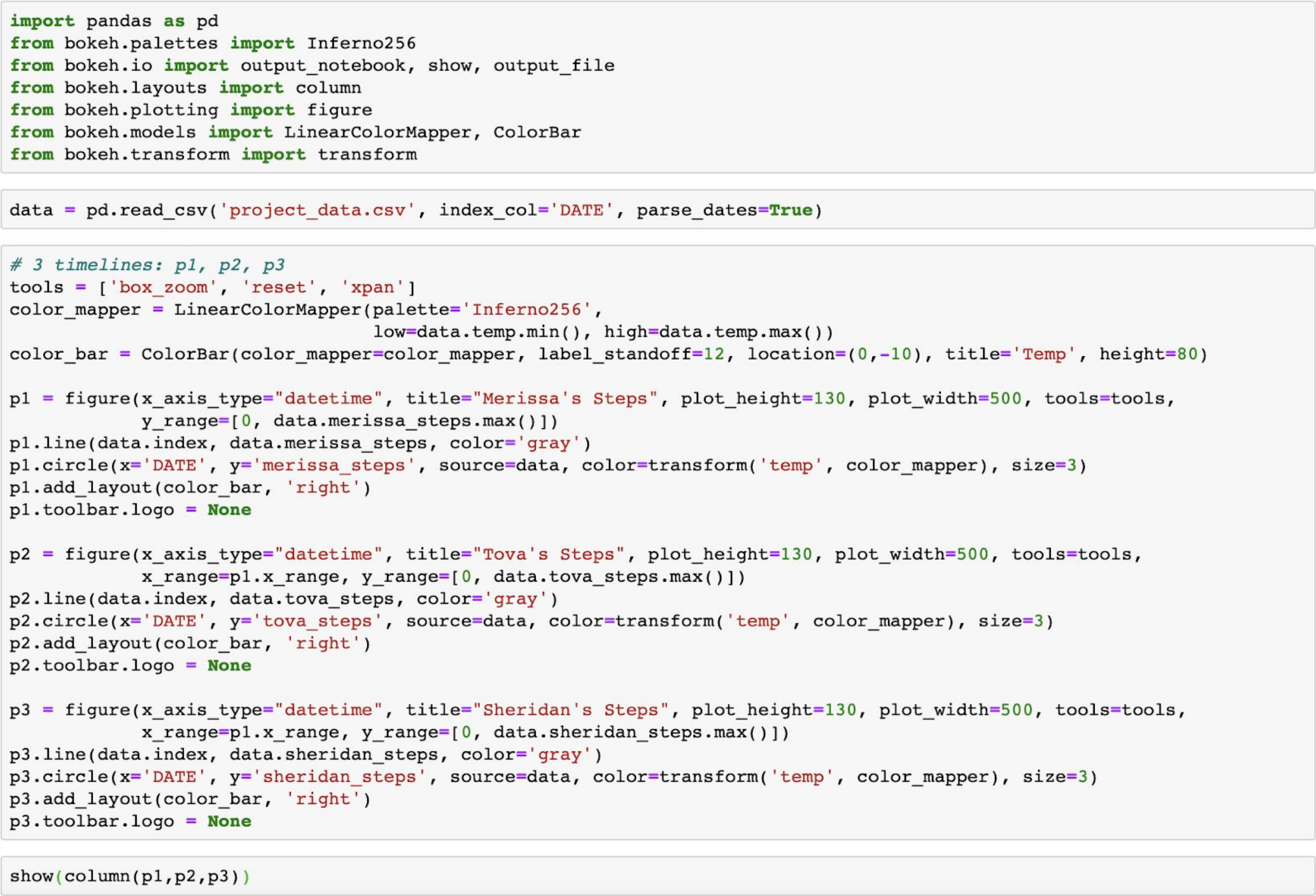
Visualization 3, by: Tova Schwartz

The third visualization was created using the Python Bokeh Library. This visualization is a time-series analysis: the number of steps taken daily (y-axis) is plotted over the course of the summer (x-axis). The interactive version of the visualization has all three scatterplots on one figure; however, for a static representation I split them up for the sake of interpretability. In the interactive version, selecting a person to study isolates that person’s data in all three visualizations. Selecting a time in one graph selects the time in the temperature - steps graphs as well. The interactions allow for synchronized selection as well as scrolling. There is a line plot imposed on the static version of the graph; this is not possible when all the graphs are combined because it creates too much noise, but in the static version it makes trends easier to analyze. The points are color-coded by temperature using the Inferno256 colormap.



The original assumption was that days with milder weather would result in a greater number of steps taken. This is because in very high temperatures people opt to stay in door or commute by car in contrast to taking a walk when the weather is beautiful. However, a cursory analysis of this visualization provides no support for said hypothesis; this is true for the data as a whole as well as for each individual's data separately. This is confirmed by the regression of temperature against steps.

(Code for the visualizations shown above.)



```python
import pandas as pd
from bokeh.palettes import Inferno256
from bokeh.io import output_notebook, show, output_file
from bokeh.layouts import column
from bokeh.plotting import figure
from bokeh.models import LinearColorMapper, ColorBar
from bokeh.transform import transform
```

```python
data = pd.read_csv('project_data.csv', index_col='DATE', parse_dates=True)
```

```python
# 3 timelines: p1, p2, p3
tools = ['box_zoom', 'reset', 'xpan']
color_mapper = LinearColorMapper(palette='Inferno256',
                                 low=data.temp.min(), high=data.temp.max())
color_bar = ColorBar(color_mapper=color_mapper, label_standoff=12, location=(0,-10), title='Temp', height=80)

p1 = figure(x_axis_type="datetime", title="Merissa's Steps", plot_height=130, plot_width=500, tools=tools,
            y_range=[0, data.merissa_steps.max()])
p1.line(data.index, data.merissa_steps, color='gray')
p1.circle(x='DATE', y='merissa_steps', source=data, color=transform('temp', color_mapper), size=3)
p1.add_layout(color_bar, 'right')
p1.toolbar.logo = None

p2 = figure(x_axis_type="datetime", title="Tova's Steps", plot_height=130, plot_width=500, tools=tools,
            x_range=p1.x_range, y_range=[0, data.tova_steps.max()])
p2.line(data.index, data.tova_steps, color='gray')
p2.circle(x='DATE', y='tova_steps', source=data, color=transform('temp', color_mapper), size=3)
p2.add_layout(color_bar, 'right')
p2.toolbar.logo = None

p3 = figure(x_axis_type="datetime", title="Sheridan's Steps", plot_height=130, plot_width=500, tools=tools,
            x_range=p1.x_range, y_range=[0, data.sheridan_steps.max()])
p3.line(data.index, data.sheridan_steps, color='gray')
p3.circle(x='DATE', y='sheridan_steps', source=data, color=transform('temp', color_mapper), size=3)
p3.add_layout(color_bar, 'right')
p3.toolbar.logo = None
```

```python
show(column(p1,p2,p3))
```

Discussion

The very low R^2 values for each of the linear regression models tell us that temperature is not a good determinant of the number of steps walked. This is further proven when we look at the p-values associated with the coefficient on temperature, which we see are high across all models stating that temperature is not a statistically significant determinant of steps walked. The correlation coefficients are also very low across all models further proving that temperature is not a good determinant of steps walked.

We expected to see negative coefficients on temperature through the regression analysis because it is expected that when temperatures are uncomfortably hot people are less likely to be walking around and will instead seek refuge from the high temperatures. Instead, we saw that there was an overall positive relationship between temperature and steps walked. This suggests that there are other factors that affect steps walked other than (and possible more strongly than) temperature, which we were unable to account for. This is corroborated by the time series analysis that incorporates the temperature dimension. Visually this allows us to see that the fluctuation in steps is totally random and cannot be attributed to hotter days.

Model:	**Overall Steps**	**Merissa Steps**	**Sheridan Steps**	**Tova Steps**
R^2	0.005	0.020	0.001	0.019
Adjusted R^2	0.002	0.012	-0.007	0.011
Coefficient on Temperature	57.5074	131.1345	-20.0483	61.4387
P-value	0.197	0.123	0.748	0.129
Correlation Coefficient	0.0676	0.140	-0.0294	0.1380

Additionally, our small multiples visualization provided results that reject our hypothesis. The visualization shows that there is no cyclical weekly trend and activity level does not increase on weekends when there is often more free time. This can be because as students, over the summer our schedule is more flexible and does not have great rigidity. However, the results from September, when school is in session, also does not have a weekly cyclical pattern. The lack of a cyclical pattern rejects our hypothesis and raises the question of if our hectic schedules are truly responsible for inactivity if it is simply an easy scapegoat.

Conclusion

In conclusion, temperature is not significantly correlated with activity level and weekly activity does not follow a cyclical weekly pattern. Our data has proved to be highly erratic and lacking in pattern. Our analysis is not without limitations. Our analysis is done on only three subjects, over the summer, an inherently unstructured period of time, and fails to take into account several crucial variables, such as location. Since health is inherently such an important aspect of our lives, this topic demands further research. Continued research would look at the full year, and perhaps several years, and contain location data. It would also take into account the type of weather: sunny, raining, snowing. An advanced study would also control pedometer reading quality. Since not everyone has their device on them constantly the data can be missing; a possible solution is results from smartwatches and bracelets that are worn constantly. Last, and most importantly, a thorough study would require a much larger sample size. Since this analysis only included data from three individuals, it is hard to use these results to reach a general conclusion on an overall population due to the wide variation in our results. Even if the results for each individual were similar to each other, we still would not be able to use our results to reach a general conclusion as there might be gender differences (as we are all female) and other factors that affect step counts that we are unable to account for in our analysis.

References

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). “Overweight & Obesity Statistics.” *National Institute of Diabetes and Digestive and Kidney Diseases*, U.S. Department of Health and Human Services, 1 Aug. 2017, https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity.

Baskerville, R., et al. “Impact of Accelerometer and Pedometer Use on Physical Activity and Glycaemic Control in People with Type 2 Diabetes: a Systematic Review and Meta-Analysis.” *Diabetic Medicine*, vol. 34, no. 5, 2017, pp. 612–620., doi:10.1111/dme.13331.

Bumgardner, Wendy. “How Many Average Daily Steps Do People Walk?” *Verywell Fit*, Verywell Fit, 6 Dec. 2019, https://www.verywellfit.com/whats-typical-for-average-daily-steps-3435736.

Centers for Disease Control and Prevention. “National Diabetes Statistics Report.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 24 Feb. 2018, https://www.cdc.gov/diabetes/data/statistics/statistics-report.html.

Department of Health & Human Services. “Walking for Good Health.” *Better Health Channel*, Department of Health & Human Services, 30 June 2015, https://www.betterhealth.vic.gov.au/health/healthyliving/walking-for-good-health.

Le Masurier, Guy C., et al. “Pedometer-Determined Physical Activity Levels of Youth.” *Journal of Physical Activity and Health*, vol. 2, no. 2, 2005, pp. 159–168., doi:10.1123/jpah.2.2.159.

National Center for Chronic Disease Prevention and Health Promotion , Division for Heart Disease and Stroke Prevention. “Heart Disease Facts & Statistics.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 28 Nov. 2017, https://www.cdc.gov/heartdisease/facts.htm.

National Weather Service Corporate Image Web Team. “National Weather Service Climate.” *National Weather Service*, National Weather Service, 24 Oct. 2005, https://w2.weather.gov/climate/xmacis.php?wfo=okx.

Richardson, Caroline R., et al. “A Meta-Analysis of Pedometer-Based Walking Interventions and Weight Loss.” *The Annals of Family Medicine*, vol. 6, no. 1, Jan. 2008, pp. 69–77., doi:10.1370/afm.761.

Code

```
import pandas as pd
import numpy as np
import datetime

from bokeh.io import *
from bokeh.plotting import *

from bokeh.models import *
from bokeh.models.widgets import *

from bokeh.layouts import *
from bokeh.palettes import *

from bokeh.application.handlers import *
from bokeh.application import *
from bokeh.palettes import Inferno256
from bokeh.transform import transform

output_notebook()



data = pd.read_csv('project_data.csv', parse_dates=['DATE'])
print(data.head(), data.info())
available_people = ['merissa_steps', 'sheridan_steps', 'tova_steps']
data = data.melt(id_vars=['DATE', 'temp'], value_vars=available_people,
var_name='person', value_name='steps')

colors = []
for color in ['DarkCyan', 'HotPink', 'Indigo']:
    for entry in range(122):
        colors.append(color)
data['color_by_person'] = colors
data['week'] = data.DATE.dt.week


available_people = ['merissa_steps', 'sheridan_steps', 'tova_steps']

def modify_doc(doc):

    def make_dataset(person_list):
        list_of_frames = []
        for person, frame in zip(person_list, 'abc'[:len(person_list)]):
            frame = data[data['person'] == person]
            list_of_frames.append(frame)
        by_person = pd.concat(list_of_frames)

        column_for_week = pd.DataFrame()
        for week_name, data_name, num in zip(['week_' + str(week) for week
in range(23,39)],
                                             ['week_data_' + str(week) for
week in range(23,39)],
                                             list(range(23,39))):
            column_for_week[week_name] = by_person[by_person['week'] ==
num]['DATE'].values
            column_for_week[data_name] = by_person[by_person['week'] ==
num]['steps'].values

        column_for_week['color'] = by_person[by_person.week ==
23]['color_by_person'].values
        return ColumnDataSource(by_person),
ColumnDataSource(column_for_week)

    def tovas_plot(src):
        color_mapper = LinearColorMapper(palette='Inferno256',
                                 low=data.temp.min(),
high=data.temp.max())
        color_bar = ColorBar(color_mapper=color_mapper, label_standoff=12,
location=(0,-10), title='Temp', height=150)
        p1 = figure(x_axis_type="datetime", title="Everyone's Steps",
plot_height=240, plot_width=480,
                    tools=['box_select', 'reset', 'xwheel_zoom', 'xpan'])
        p1.scatter(source=src, x='DATE', y='steps',
color=transform('temp', color_mapper))
        p1.add_layout(color_bar, 'right')
        return p1

    def sheridans_plot(src):
        p2 = figure(title="Steps by temperature", plot_height=240,
plot_width=480,
                    tools=['box_select', 'reset'])
        p2.scatter(source=src, x='temp', y='steps',
line_color='color_by_person', fill_color='color_by_person')
        return p2
    def merissas_plot(src):
        weekly_plots = []
        for week_name, data_name, num in zip(['week_' + str(week) for week
in range(23,39)],
                                             ['week_data_' + str(week) for
week in range(23,39)],
                                             list(range(23,39))):
            p3 = figure(plot_width=240, plot_height=180,
                        x_axis_type='datetime',
                        x_axis_label='Day', y_axis_label='Steps',
                        title='Week {}'.format(num),
                        toolbar_location='below', tools=['hover'],
                        tooltips=[("Steps", "$y{int}")])
            p3.circle(source=src, x=week_name, y=data_name, color='color',
size=10, alpha=.5)
            weekly_plots.append(p3)
        gp = gridplot(weekly_plots, ncols=4)
        return gp



    def update(attr, old, new):
        people_to_plot = [person_selection.labels[i] for i in
                            person_selection.active]
        new_src, new_src2 = make_dataset(people_to_plot)

        src.data.update(new_src.data)
        src2.data.update(new_src2.data)



    person_selection = CheckboxGroup(labels=available_people, active = [0,
1, 2])
    person_selection.on_change('active', update)

    controls = WidgetBox(person_selection)

    initial_people = [person_selection.labels[i] for i in
person_selection.active]

    src, src2 = make_dataset(initial_people)

    p1 = tovas_plot(src)
    p2 = sheridans_plot(src)
    gp = merissas_plot(src2)
    layout = column(controls, row(p1, p2), gp)
    doc.add_root(layout)

handler = FunctionHandler(modify_doc)
app = Application(handler)

   show(app)
```