0.1 1.b. Map traffic speed to Google Plus Codes

Google Plus Codes divide up the world uniformly into rectangular slices (link). Let's use this to segment traffic speeds spatially. Take a moment to answer: **Is this spatial structure effective for summarizing traffic speed?** Before completing this section, substantiate your answer with examples of your expectations (e.g., we expect A to be separated from B). After completing this section, substantiate your answer with observations you've made.

No, while this spatial structure is very effective for giving a universal street address for people or places that don't have one, it lacks in its ability to differentiate between traffic speeds of adjacent roads that would be located within the same small grid. For example let's say we are observing a grid that goes from one exit of a freeway to the next. The grid system would have difficulty determining the difference between the north and south bound traffic speed, if northbound has no traffic and southbound has lots of traffic but they lie within the same grid, how would this spatial structure determine this difference?

0.1.1 1.b.v. How well do plus code regions summarize movement speeds?

The following will give us an idea of how well the average represents traffic speed per plus code region. For these questions, we'll refer to a "plus code region" as a "cluster":

1. **Plot a histogram of the within-cluster standard deviation**.
2. **Compute across-cluster average of within-cluster standard deviation**.
3. **Compute across-cluster standard deviation of within-cluster average speeds**.
4. **Is this average variance reasonable?** To assess what "reasonable" means, consider these questions and how to answer them: (1) Do plus codes capture meaningful subpopulations? (2) Do differences between subpopulations outweigh differences within a subpopulation? Use the statistics above to answer these questions, and compute any additional statistics you need. Additionally explain *why these questions are important to assessing the quality of a spatial clustering.*

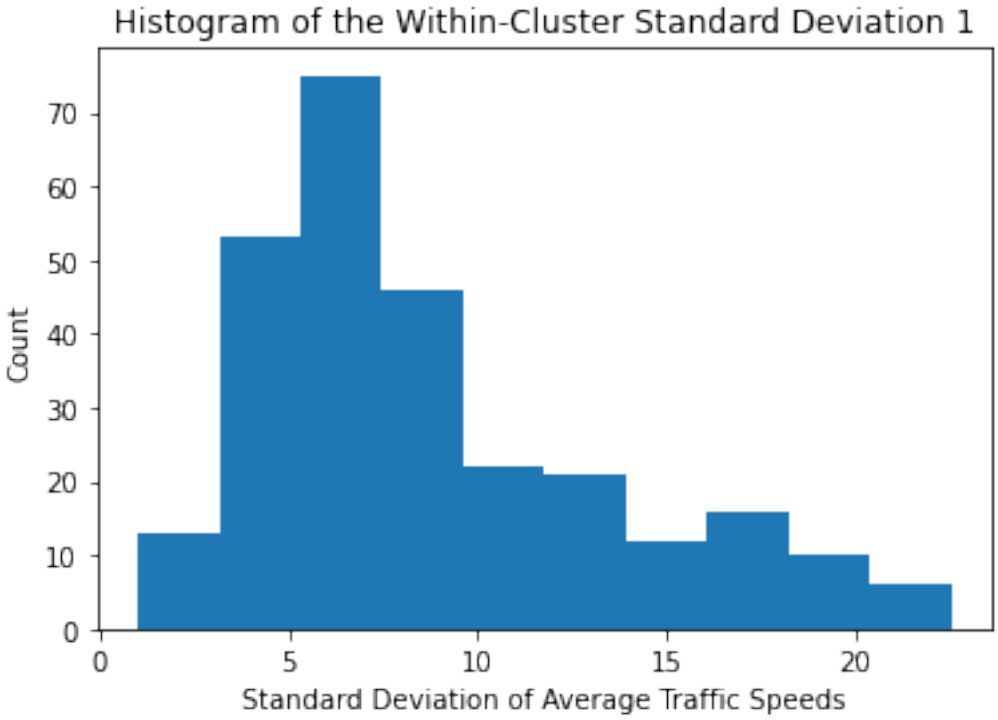
Hint: Run the autograder first to ensure your variance average and average variance are correct, before starting to draw conclusions.

In the first cell, write your written answers. In the second cell, complete the code.

While the concept of using plus codes does not perfectly capture the intricate differences in traffic speeds, it again consistently comes within a ten mile per hour accuracy for the true speed of traffic. Plus codes do indeed capture meaningful subpopulations and it does so in a manner that is easy to understand as well as adjust if given more data. The differences between subpopulations do outweigh the differences within a subpopulation, because at any given point on a road there is usually at least a difference of ten miles per hour between the slowest and fastest cars. So given the fact that the average standard deviation is only around 9.4 mph, the average variance is indeed reasonable. This implementation of plus codes allows data scientists to create meaningful subpopulations that create a reasonable average variance, and could be furthered by cross-referencing current traffic speed data collected on other platforms such as Waze or Google Maps. While in further exploration of this data we can make these plus codes more specific in order to reduce this average variance.

In [16]:
```
data_grouped = speeds_to_gps.groupby(["plus_latitude_idx", "plus_longitude_idx"])

# compute traffic speed variance in each plus code region
speed_variance_by_pluscode = data_grouped.var()["speed_mph_mean"]
# plot a histogram
average_variance_by_pluscode = data_grouped.std()["speed_mph_mean"].mean()
variance_average_by_pluscode = data_grouped.mean()["speed_mph_mean"].std()

variance_average_by_pluscode

data = speeds_to_gps.groupby(["plus_latitude_idx", "plus_longitude_idx"]).std()['speed_mph_mean']

plt.hist(data)
plt.ylabel('Count')
plt.xlabel('Standard Deviation of Average Traffic Speeds')
plt.title('Histogram of the Within-Cluster Standard Deviation 1')
```

Out[16]:
```
Text(0.5, 1.0, 'Histogram of the Within-Cluster Standard Deviation 1')
```



0.1.2 1.c.iv. How well do census tracts summarize movement speeds?

The following will give us an idea of how well the average represents traffic speed per plus code region. For these questions, we'll refer to a "census tract" as a "cluster":

1. **Plot a histogram of the within-cluster standard deviation**.
2. **Compute across-cluster average of within-cluster standard deviation**.
3. **Compute across-cluster standard deviation of within-cluster average speeds**.
4. **Is this average variance reasonable?** To assess what "reasonable" means, consider these questions and how to answer them: (1) Do plus codes capture meaningful subpopulations? (2) Do differences between subpopulations outweigh differences within a subpopulation? Use these ideas to assess whether the average standard deviation is high or not.

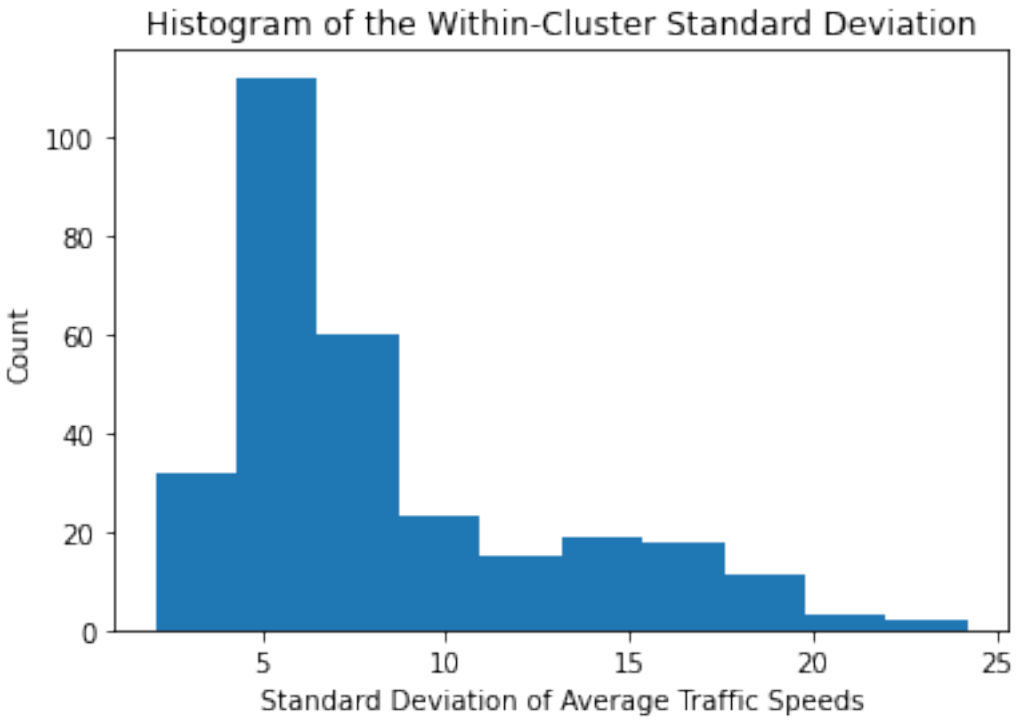
Note: We are using the speed metric of miles per hour here.

Just like before, please written answers in the first cell and coding answers in the second cell.

While the concept of using plus codes still does not perfectly capture the intricate differences in traffic speeds, it again consistently comes within a ten mile per hour accuracy for the true speed of traffic. Plus codes do indeed capture meaningful subpopulations and it does so in a manner that is easy to understand as well as adjust if given more data. The differences between subpopulations do outweigh the differences within a subpopulation, because at any given point on a road there is usually at least a difference of ten miles per hour between the slowest and fastest cars. So given the fact that the average standard deviation is only around 8.3 mph, the average variance is indeed reasonable and it has decreased in average variance from our similar assessment in 1.b.v. This implementation of plus codes allows data scientists to create meaningful subpopulations that create a reasonable average variance, and could be furthered by cross-referencing current traffic speed data collected on other platforms such as Waze or Google Maps. Comparing the first and second histograms created we can observe that there is a slight lower density of these standard deviations, and while there is still a right tail it is decreasing in size.

In [25]:
```
speed_variance_by_tract = speeds_by_tract.var()["speed_mph_mean"]
average_variance_by_tract = speeds_by_tract.std()["speed_mph_mean"].mean()
variance_average_by_tract = speeds_by_tract.mean()["speed_mph_mean"].std()

data = speeds_by_tract.std()['speed_mph_mean']

plt.hist(data)
plt.ylabel('Count')
plt.xlabel('Standard Deviation of Average Traffic Speeds')
plt.title('Histogram of the Within-Cluster Standard Deviation')
```

Out[25]:
```
Text(0.5, 1.0, 'Histogram of the Within-Cluster Standard Deviation')
```



0.2 1.d. What would be the ideal spatial clustering?

This is an active research problem in many spatiotemporal modeling communities, and there is no single agreed-upon answer. Answer both of the following specifically knowing that you'll need to analyze traffic patterns according to this spatial clustering:

1. **What is a good metric for a spatial structure?** How do we define good? Bad? What information do we expect a spatial structure to yield? Use the above parts and questions to help answer this.
2. **What would you do to optimize your own metric for success in a spatial structure?**

See related articles:

- Uber's H3 link, which divides the world into hexagons
- Traffic Analysis Zones (TAZ) link, which takes census data and additionally accounts for vehicles per household when dividing space

1. Uber's H3 Hexagonal Hierarchical spatial indexing was based around the idea that people are always in motion and in order to capture that motion to efficiently optimize ride pricing and dispatch, and to visualize and explore spatial data we need to minimize the quantization error introduced when users move through a city. A 'good metric for spatial structuring' has been defined as one that gives the most accurate socio-economic data. We expect a spatial structure to yield many things including: the size of the individual bin (or in Uber's case, hexagon) in relation to the size of the greater surrounding area (ie block as part of a city), the population within that bin, and the speed of the traffic within that bin. A spatial structure that can do these tasks accurately is typically defined as 'good'.

2. Our metric for success is primarily based upon accurately predicting traffic speed in each bin, as a part of the greater surrounding area. So similar to the approach that Uber has taken, we are also looking to minimize the quantization error introduced when users move through a city. To effectively achieve this goal we have binned the world into squares based on latitude and longitude degrees. As the bins get smaller there is greater computations required but the traffic predictions do reduce in average variance.

0.2.1 2.a.i. Sort census tracts by average speed, pre-lockdown.

Consider the pre-lockdown period to be March 1 - 13, before the first COVID-related restrictions (travel bans) were announced on March 14, 2020.

1. **Report a DataFrame which includes the *names* of the 10 census tracts with the lowest average speed**, along with the average speed for each tract.
2. **Report a DataFrame which includes the *names* of the 10 census tracts with the highest average speed**, along with the average speed for each tract.
3. Do these names match your expectations for low speed or high speed traffic pre-lockdown? What relationships do you notice? (What do the low-speed areas have in common? The high-speed areas?) For this specific question, answer qualitatively. No need to quantify. **Hint**: Look up some of the names on a map, to understand where they are.
4. **Plot a histogram for all average speeds, pre-lockdown**.
5. You will notice a long tail distribution of high speed traffic. What do you think this corresponds to in San Francisco? Write down your hypothesis.

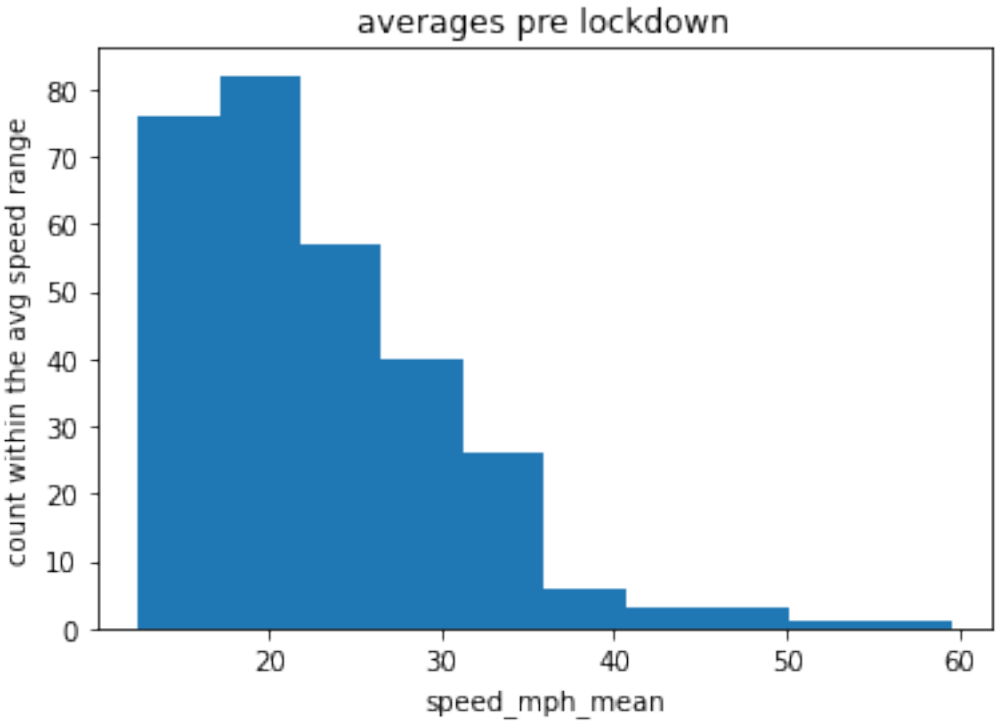
Hint: To start off, think about what joins may be useful to get the desired DataFrame.

For the top 10 lowest average census tracts speeds, I would have to assume that they would all be in the heart of San Francisco as that is the most densely populated and heavy traffic sites in the county, thus these spots would have low average traffic speeds. Whereas for the top 10 highest average census tracts speeds, I would have to assume that they would be on the edge of the city, the farthest points from the center of the city since those places are less densely populated and have on average less traffic, thus the average traffic speeds would be higher. As for the tail, I think that there would likely be a right tail of high traffic speed points that would be represented by these areas on the edge of the city or on the less packed freeways far from the city center.

Plot the histogram

```
In [50]: plt.title("averages pre lockdown")
         plt.hist(averages_pre)
         plt.xlabel("speed_mph_mean")
         plt.ylabel("count within the avg speed range")
```

```
Out[50]: Text(0, 0.5, 'count within the avg speed range')
```



0.2.2 2.a.ii. Sort census tracts by average speed, post-lockdown.

I suggest checking the top 10 and bottom 10 tracts by average speed, post-lockdown. Consider the post-lockdown period to be March 14 - 31, after the first COVID restrictions were established on March 14, 2020. It's a healthy sanity check. For this question, you should report:

- **Plot a histogram for all average speeds, post-lockdown.**
- **What are the major differences between this post-lockdown histogram relative to the pre-lockdown histogram above**? Anything surprising? What did you expect, and what did you find?

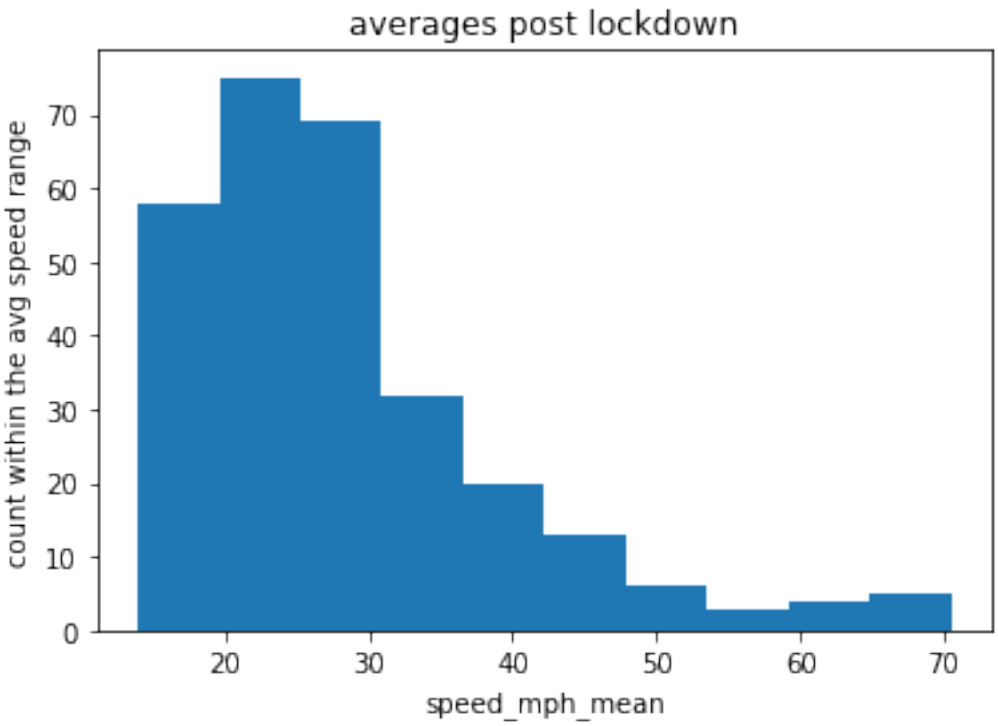
Write the written answers in the cell below, and the coding answers in the cells after that.

The major differences between this post-lockdown histogram relative to the pre-lockdown histogram are the exact differences that we would expect: a higher distribution of faster traffic speeds in the San Francisco area, but no changes that are extraordinary outstanding since the city still is confined to small intricate blocks with many stop signs and street lights. Post-lockdown there are many fewer people on the roads and thus we would expect traffic speeds to be increased by the decrease in cars and congestion; and comparing the two histograms that is indeed what we can observe.

Plot the histogram

```
In [53]: plt.hist(averages_post)
         plt.title("averages post lockdown")
         plt.xlabel("speed_mph_mean")
         plt.ylabel("count within the avg speed range")
```

```
Out[53]: Text(0, 0.5, 'count within the avg speed range')
```

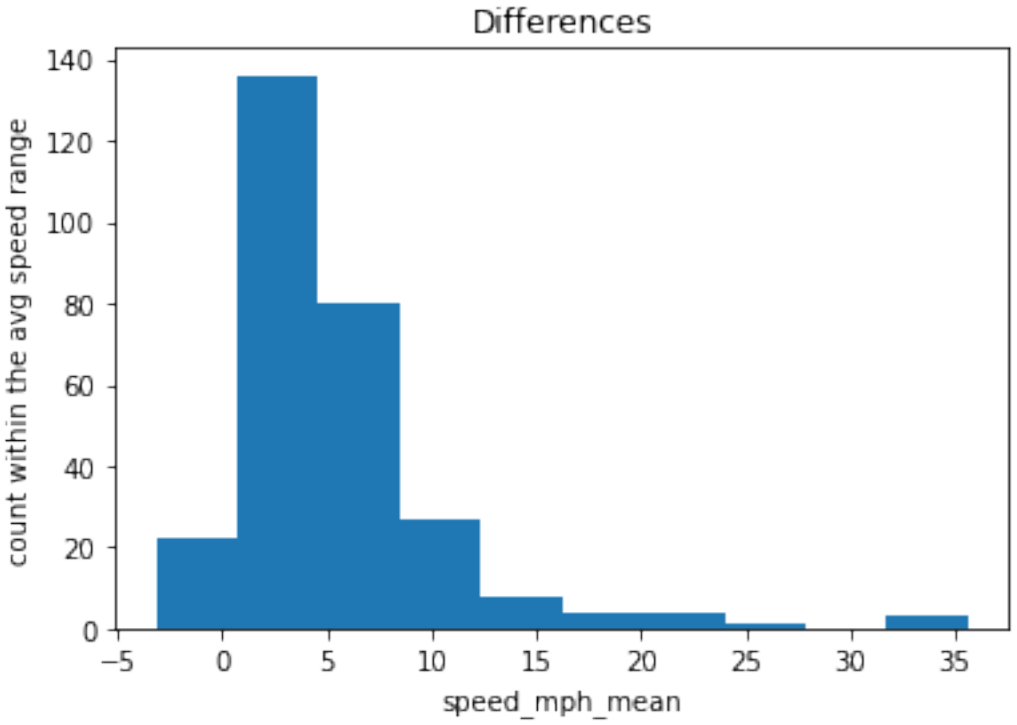


0.2.3 2.a.iii. Sort census tracts by change in traffic speed from pre to post lockdown.

For each segment, compute the difference between the pre-lockdown average speed (March 1 - 13) and the post-lockdown average speed (March 14 - 31). **Plot a histogram of all differences.** Sanity check that the below histogram matches your observations of the histograms above, on your own.

In [54]:
```
# The autograder expects differences to be a series object with index
# MOVEMENT_ID.
differences = averages_post - averages_pre
# plot the differences
plt.hist(differences)
plt.title("Differences")
plt.xlabel("speed_mph_mean")
plt.ylabel("count within the avg speed range")
```

Out[54]:
```
Text(0, 0.5, 'count within the avg speed range')
```



In [55]:
```
grader.check("q2aiii")
```

```
Out[55]: q2aiii results: All test cases passed!
```

0.2.4 2.a.iv. Quantify the impact of lockdown on average speeds.

1. **Plot the average speed by day, across all segments**. Be careful not to plot the average of census tract averages instead. Recall the definition of segments from Q1.
2. Is the change in speed smooth and gradually increasing? Or increasing sharply? Why? Use your real-world knowledge of announcements and measures during that time, in your explanation. You can use this list of bay area COVID-related dataes: https://abc7news.com/timeline-of-coronavirus-us-covid-19-bay-area-sf/6047519/

In [56]:
```
# Autograder expects this to be a series object containing the
# data for your line plot -- average speeds per day.
speeds_daily = speeds_to_tract.groupby("day").agg(np.mean)["speed_mph_mean"]
speeds_daily
```

Out[56]:
```
day
1     23.633128
2     22.324856
3     21.826422
4     21.931746
5     21.985687
6     22.100850
7     22.851309
8     24.049457
9     23.008859
10    22.661657
11    22.776953
12    23.078484
13    23.320432
14    24.177464
15    25.698629
16    24.771821
17    28.912335
18    28.811352
19    29.502508
20    28.939733
21    30.062931
22    30.766490
23    30.236890
24    30.120710
25    30.251735
26    30.136657
27    30.344251
28    30.489458
29    31.062707
30    30.440111
31    30.870653
Name: speed_mph_mean, dtype: float64
```

Write your written answer in the cell below

The change in speed is gradually increasing; for the first two weeks, the speeds mostly stayed the same but by the third week we see a consistent gradual increase. In the first week we can actually observe an decrease in traffic speed likely because most people were still going into work, living their normal lives, and also because there was some panic from people who thought the world was going to end and they needed to stockpile, like those who desperately needed toilet paper. For these reasons we can see a slight decrease in traffic speed during the first week, and in the second week we see an increase to around the average speed pre-lockdown, likely because people have stopped going out as much but are still fearful and going out to prepare for a long lockdown. Finally after the first two weeks the gradual increase begins and people no longer need to stockpile and have made the adjustment to working from home.

0.2.5 2.a.v. Quantify the impact of pre-lockdown average speed on change in speed.

1. Compute the correlation between change in speed and the *pre*-lockdown average speeds. Do we expect a positive or negative correlation, given our analysis above?
2. Compute the correlation between change in speed and the post-lockdown average speeds.
3. **How does the correlation in Q1 compare with the correlation in Q2?** You should expect a significant change in correlation value. What insight does this provide about traffic?

Written answers in the first cell, coding answerts in the following cell.

The correlation post-lockdown had a much stronger positive linear relationship when compared to correlation pre-lockdown. We observed that the correlation difference pre-lockdown had a coefficient of 0.46 whereas the correlation difference post-lockdown had a coefficient of 0.79, representing a stronger positive linear association. This insight allows us to predict that traffic will be more affected post-lockdown by other large changes in society. Therefore there will be more change in speeds going forward in our post-lockdown society and thus traffic speeds will likely decrease as a result of increasing changes in speeds (greater fluctuation of traffic speed often results in more traffic, or slower traffic speeds overall).

0.2.6 2.b.i. Visualize spatial heatmap of average traffic speed per census tract, pre-lockdown.

Visualize a spatial heatmap of the grouped average daily speeds per census tract, which you computed in previous parts. Use the geopandas chloropleth maps. **Write your observations, using your visualization, noting down at least 2 areas or patterns of interest**. These may be a local extrema, or a region that is strangely all similar.

Hint: Use `to_crs` and make sure the `epsg` is using the Pseudo-Mercator projection.

Hint: You can use `contextily` to superimpose your chloropleth map on a real geographic map.

Hint You can set a lower opacity for your chloropleth map, to see what's underneath, but be aware that if you plot with too low of an opacity, the map underneath will perturb your chloropleth and meddle with your conclusions.

Written answers in the first cell, coding answers in the second cell.

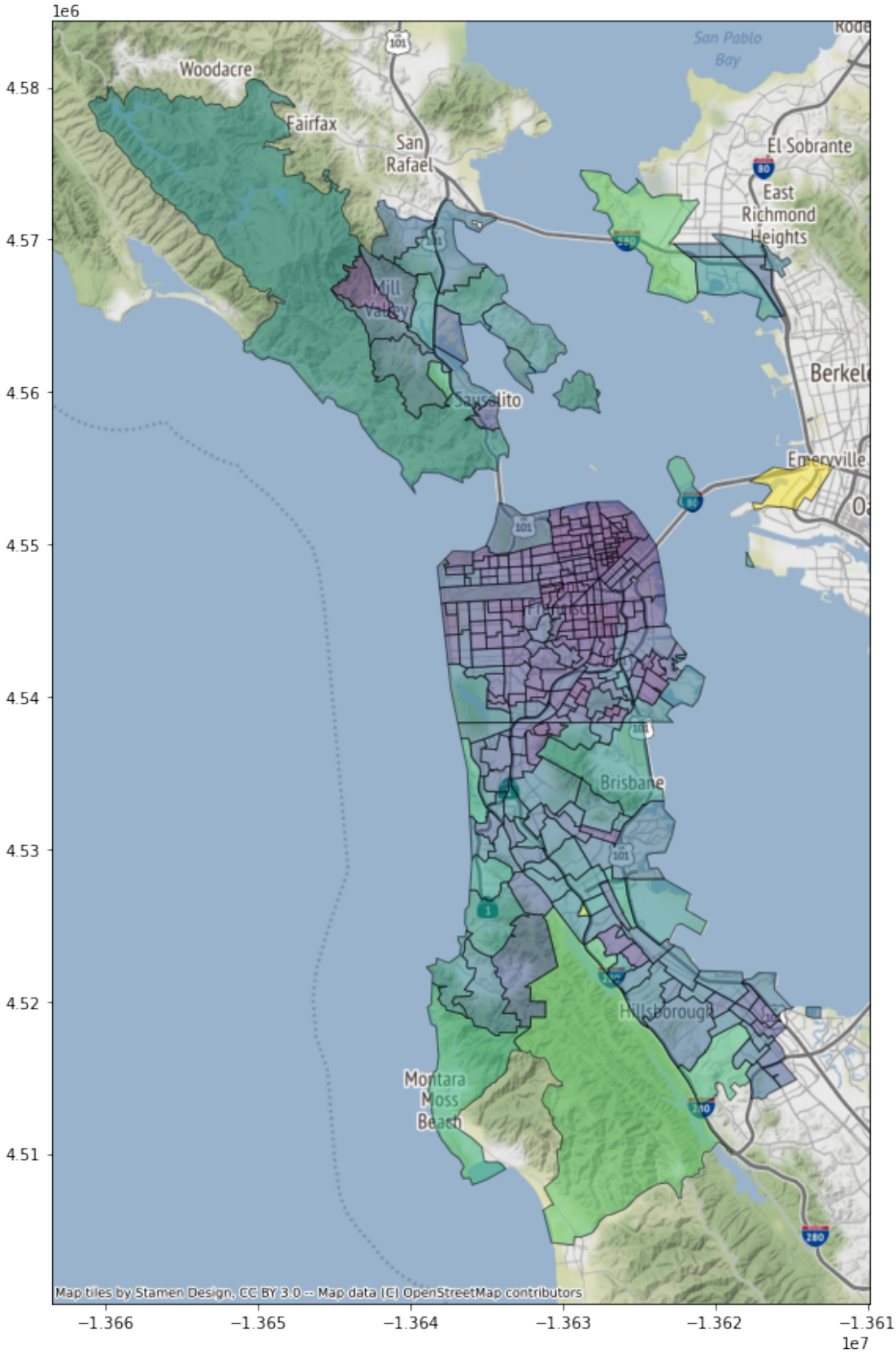
-Local extrema in the heart of the city of San Francisco near Embarcadero

-Local extrema in north bay on the northeast section of mill valley

-The only section that is yellow is the southwest section of Emeryville near the Bay Bridge exit/entrance, likely means that this area is very traffic dependent on time of day

-Local extrema near San Mateo

In [61]:
```
gpd_pre = gpd.GeoDataFrame(averages_pre_named)
ax = gpd_pre.to_crs(3857).plot(column = "speed_mph_mean", figsize=(15, 15),alpha=0.5, edgecolo
cx.add_basemap(ax)
```


1e6
4.58
4.57
4.56
4.55
4.54
4.53
4.52
4.51
Woodacre
Fairfax
San Rafael
San Pablo Bay
El Sobrante
East Richmond Heights
Mill Valley
Sausalito
Berkele
Emeryville
Brisbane
Hillsborough
Montara
Moss Beach
Map tiles by Stamen Design, CC BY 3.0 — Map data (C) OpenStreetMap contributors
−1.366
−1.365
−1.364
−1.363
−1.362
−1.361
1e7

0.2.7 2.b.ii. Visualize change in average daily speeds pre vs. post lockdown.

Visualize a spatial heatmap of the census tract differences in average speeds, that we computed in a previous part. **Write your observations, using your visualization, noting down at least 2 areas or patterns of interest.** Some possible ideas for interesting notes: Which areas saw the most change in average speed? Which areas weren't affected? Why did some areas see *reduced* average speed?

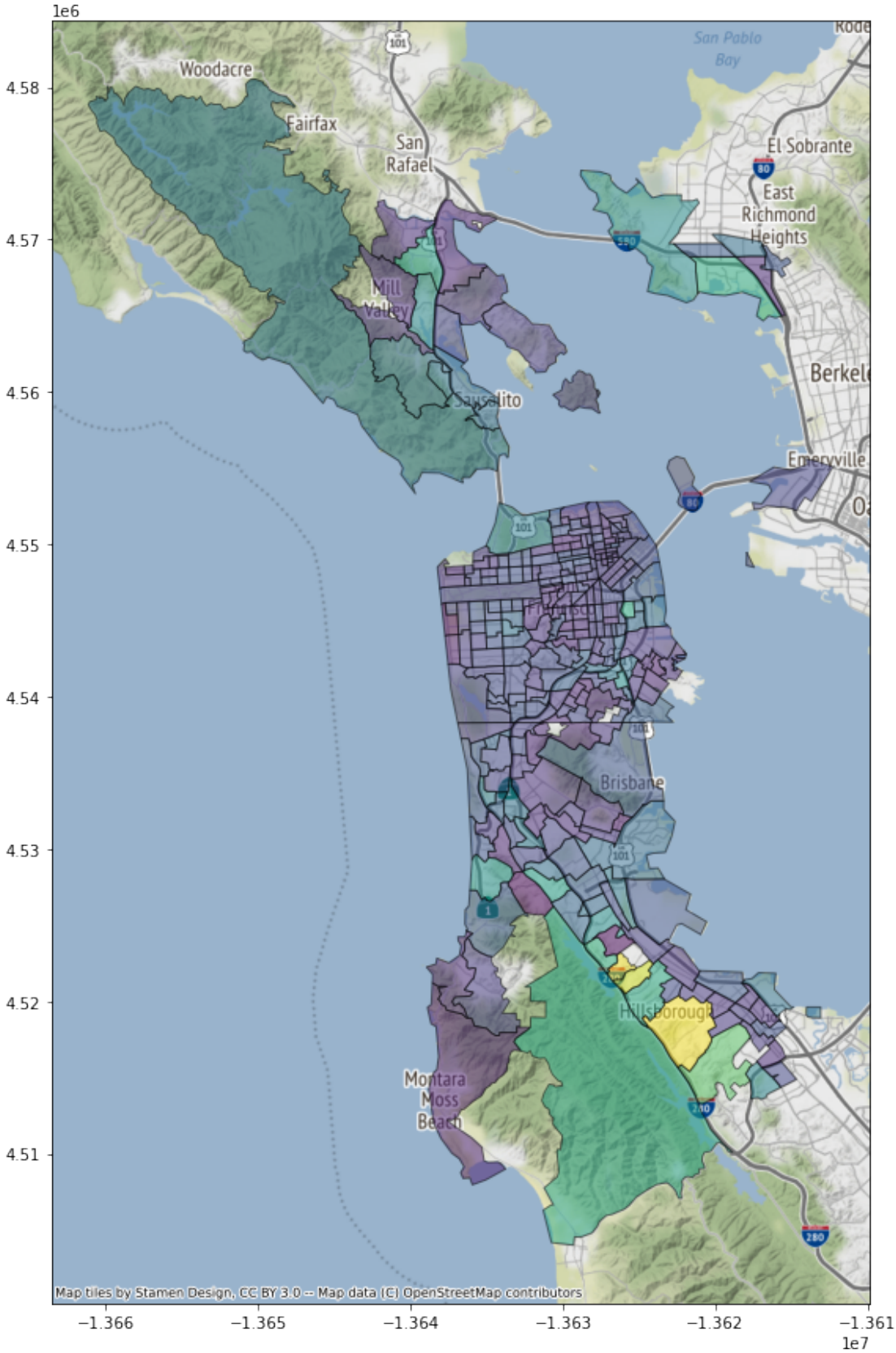
First cell is for the written answers, second cell is for the coding answers.

-The west coastal region near Montara Moss beach saw a large change in average speed. It is likely that many of the wealthy people from San Francisco retired to their second homes, or away from their city apartments, to the beach to get away from densely populated areas.

-Similarly with Mill Valley and south San Rafael, there was a large change in average speed because people wanted to get out of the dense city of SF and away from people.

-Southwest Emeryville as well changed from yellow to purple, representing that there was also a large change in traffic speed, likely again people moving away from the two major cities it lies between.

In [62]:
```
averages_pre_named["difference_mpg_means"] = differences
gpd_diff = gpd.GeoDataFrame(averages_pre_named)
ax_diff = gpd_pre.to_crs(3857).plot(column = "difference_mpg_means",figsize=(15, 15), alpha=0.5
cx.add_basemap(ax_diff)
```


1e6
4.58
4.57
4.56
4.55
4.54
4.53
4.52
4.51
Woodacre
Fairfax
San Rafael
San Pablo Bay
El Sobrante
East Richmond Heights
Mill Valley
Sausalito
Berkele
Emeryville
Brisbane
Hillsborough
Montara
Moss Beach
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
−1.366
−1.365
−1.364
−1.363
−1.362
−1.361
1e7