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.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

```
In [ ]: speed_variance_by_pluscode = speeds_to_gps.groupby(["plus_latitude_idx","plus_longitude_idx"]).s
        plt.hist(speed_variance_by_pluscode);
        plt.xlabel("Within-Cluster Standard Deviation of Average Speed")
        plt.ylabel("Number of Clusters")
        plt.title("Distribution of Average Speed Standard Deviations Amongst Clusters")
        # plot a histogram
        average_variance_by_pluscode = speed_variance_by_pluscode.mean()
        variance_average_by_pluscode = speeds_to_gps.groupby(["plus_latitude_idx","plus_longitude_idx"]]
```

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.

Type your answer here, replacing this text.

```
In [ ]: speed_variance_by_tract = speeds_by_tract.std()["speed_mph_mean"]
        plt.hist(speed_variance_by_tract)
        plt.xlabel("Within-Cluster Standard Deviation of Average Speed")
        plt.ylabel("Number of Clusters")
        plt.title("Distribution of Average Speed Standard Deviations Amongst Clusters")
        average_variance_by_tract = speed_variance_by_tract.mean()
        variance_average_by_tract = speeds_by_tract.mean()["speed_mph_mean"].std()
```

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

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

Plot the histogram

```
In [ ]:
```

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.

Type your answer here, replacing this text.

Plot the histogram

```
In [ ]:
```

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 [ ]: # The autograder expects differences to be a series object with index
        # MOVEMENT_ID.
        differences = averages_post_named.set_index('DISPLAY_NAME').subtract(averages_pre_named.set_inde
        # plot the differences
        plt.hist(differences);
        plt.xlabel("Within-Cluster Difference in Average Speeds");
        plt.ylabel("Number of Clusters");
        plt.title("Distribution of the Difference in Average Speeds of Clusters Pre-Lockdown and Post-Lo
```

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

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 [ ]: # 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").mean()["speed_mph_mean"]
        plt.plot(speeds_daily);
        plt.xlabel("Day");
        plt.ylabel("Average Speed");
        plt.title("Average Speed Across All Segments in March 2020");
```

Write your written answer in the cell below

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

```
In [ ]: speeds_to_tract_cond = speeds_to_tract[["MOVEMENT_ID", "geometry", "speed_mph_mean"]]

        #heatmap = speeds_to_tract_cond.to_crs(epsg=3857).dissolve(by="MOVEMENT_ID", aggfunc='mean').pl
        #cx.add_basemap(heatmap)
```

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.

Type your answer here, replacing this text.

```
In [ ]: speeds_to_tract_pre = speeds_to_tract[speeds_to_tract["day"] < 14]
        speeds_to_tract_post = speeds_to_tract[speeds_to_tract["day"] >= 14]
        #poly_avg_pre = speeds_to_tract_pre.to_crs(epsg=3857).dissolve(by="MOVEMENT_ID", aggfunc="mean")
        #poly_avg_post = speeds_to_tract_post.to_crs(epsg=3857).dissolve(by="MOVEMENT_ID", aggfunc="mea
        #speeds_to_tract_cond2 = speeds_to_tract_cond.to_crs(epsg=3857).dissolve(by="MOVEMENT_ID", aggf
        #speeds_to_tract_cond2["diff"] = poly_avg_post["speed_mph_mean"].subtract(poly_avg_pre["speed_m
        #heatmap2 = speeds_to_tract_cond2.plot(column='diff', figsize=(10, 10), alpha=0.5, edgecolor='k
        #cx.add_basemap(heatmap2)
```

0.2.8 4.a.ii. Train and evaluate linear model on pre-lockdown data.

1. **Train a linear model that forecasts the next day's speed average** using your training dataset `X_train`, `y_train`. Specifically, predict $y_{(i,t)}$ from $X_{(i,t)}$, where

- $y_{(i,t)}$ is the daily speed average for day t and census tract i
- $X_{(i,t)}$ is a vector of daily speed averages for days $t-5, t-4, t-3, t-2, t-1$ for census tract i

2. **Evaluate your model** on your validation dataset `X_val`, `y_val`.
3. **Make a scatter plot**, plotting predicted averages against ground truth averages. Note the perfect model would line up all points along the line $y = x$.

Our model is quantitatively and qualitatively pretty accurate at this point, training and evaluating on pre-lockdown data.

```
In [ ]: reg = LinearRegression().fit(X_train, y_train) # set to trained linear model
        score = reg.score(X_val, y_val) # report r^2 score

        # create the scatter plot below
        plt.scatter(reg.predict(X_train), y_train);
        plt.ylabel("Ground Truth Averages");
        plt.xlabel("Predicted Averages");
        plt.title("Predicted Averages vs. Ground Truth Averages on Pre-Lockdown Data");
```

Make scatter plot below.

```
In [ ]: plt.scatter(reg.predict(x_pre), y_post);
        plt.ylabel("Ground Truth Averages");
        plt.xlabel("Predicted Averages");
        plt.title("Predicted Averages vs. Ground Truth Averages on Pre-Lockdown Data");
```

0.2.9 4.b.ii. Report model performance temporally

1. **Make a line plot** showing performance of the original model throughout all of March 2020.
2. **Report the lowest point on the line plot**, reflecting the lowest model performance.
3. **Why is model performance the worst on the 17th?** Why does it begin to worsen on march 15th? And continue to worsen? Use what you know about covid measures on those dates. You may find this webpage useful: https://abc7news.com/timeline-of-coronavirus-us-covid-19-bay-area-sf/6047519/
4. **Is the dip in performance on the 9th foreshadowed** by any of our EDA?
5. **How does the model miraculously recover on its own?**
6. **Make a scatter plot**, plotting predicted averages against ground truth averages *for model predictions on March 17th.* Note the perfect model would line up all points along the line $y = x$. When compared against previous plots of this nature, this plot looks substantially worse, with points straying far from $y = x$.

Note: Answer questions 2-5 in the Markdown cell below. Q1 and Q6 are answered in the two code cells below.

Type your answer here, replacing this text.

Generate line plot.

Generate a scatter plot.

```
In [ ]: ...
```

0.2.10 4.c.i. Learn delta off of a moving bias

According to our previous work in EDA, the average speed shoots upwards sharply. As a result, our trick to learn delta the around the average and to naively assume that the average of day t is the average for day $t + 1$. We will do this in 4 steps:

1. **Create a dataset for your delta model**.
2. **Train your delta model** on pre-lockdown data.
3. **Evaluate your model on pre-lockdown data**, to ensure that the model has learned to a satisfactory degree, in the nominal case. Remember the naive model achieved 0.97 r^2 on pre-lockdown data.
4. **Evaluate your model on the 17th**, to compare against the naive model also evaluated on that day. Notice that your r^2 score has improved by 10%+. Why is your delta model so effective for the 17th?
5. **Evaluate your model on the 14th**, to compare against the naive model also evaluated on that day. Notice that your r^2 score is now complete garbage. Why is your delta so ineffective for the 14th?

Hint: As you build your datasets, always check to make sure you're using the right days! It's easy to have a one-off error that throws off your results.

Write your written questions in the next cell, then write the code in the following cells.

Type your answer here, replacing this text.