Assignment: Exploratory data analysis

TODO: Edit this cell to fill in your NYU Net ID and your name:

- **Net ID**:
- **Name**:

Introduction

In this assignment, we will practice using exploratory data analysis on Google's COVID-19 Community Mobility data.

This data was collected from Google Maps users around the world over the last few months - including you, *if* you have Google Maps on your phone and have turned on the Location History setting. It combines location history from a large number of users to capture the overall increase or decrease in time spent in places such as: retail and recreation facilities, groceries and pharmacies, parks, transit stations, workplaces, and residences.

As you work through this notebook, you will see that some text and code cells are marked with a "TODO" at the top. You'll have to edit these cells to fill in the code or answer the questions as indicated.

Learn about the data

First, it is worthwhile to learn more about the data: how it is collected, what is included, how Google gets consent to collect this data, and how user privacy is protected. Google provides several resources for learning about the data:

- Blog post
- About this data
- Understand the data

Read in data

Now you are ready to read the data into your notebook.

Visit Google's web page for the COVID-19 Community Mobility project to get the URL for the data.

(Specific instructions will depend on your browser and operating system, but on my laptop, I can get the URL by right-clicking on the button that says "Download global CSV" and choosing "Copy Link Address".)

Then, in the following cells, use that URL to read the data into a pandas Data Frame called `df`. (You can follow the example in the "Exploratory data analysis" notebook from this week's lesson.)

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# TODO Q1
# url = ...
# df = ...
```

Use the `info()` and `head()` functions to show some basic information about the data and to look at the first few samples.

```
# TODO Q2
# use info()
# TODO Q3
# use head()
```

Basic data manipulations

The data includes a date field, but it may have been read in as a string, rather than as a `datetime`. If that's the case, use `to_datetime()` to convert the field into a datetime format. (You can follow the example in the "Exploratory data analysis" notebook from this week's lesson.)

Then, use `info()` again to make sure your change was applied. Note the difference in the output, relative to the cell above.

```
# TODO Q4
# df['date'] = ...
```

Next, you are going to extract the subset of data for the location of your choice. You can choose any location *except* Brooklyn, New York. (You can't choose Brooklyn because the example code I'm about to show you is for Brooklyn.)

The data is reported for different regions, with different levels of granularity available. This is best explained by example:

Suppose I want the overall trend from the entire U.S. I would use the subset of data where `country_region` is equal to "United States" and `sub_region_1` is null:

```
df_subset = df[(df['country_region'].eq("United States")) & (df['sub_region_1'].isnull())]
```

Suppose I want the overall trend from the entire state of New York: I would use the subset of data where `country_region` is equal to "United States", `sub_region_1` is equal to "New York", and `sub_region_2` is null:

```
df_subset = df[(df['country_region'].eq("United States")) & (df['sub_region_1'].eq("New
    York")) & (df['sub_region_2'].isnull())]
```

Suppose I want the overall trend from Brooklyn, New York (Kings County): I would use the subset of data where `country_region` is equal to "United States", `sub_region_1` is equal to "New York", and `sub_region_2` is equal to "Kings County":

```
df_subset = df[(df['country_region'].eq("United States")) & (df['sub_region_1'].eq("New
    York")) & (df['sub_region_2'].eq("Kings County"))]
```

In the following cell(s), fill in the code to create a data frame `df_subset` with data from a single location. You can go down to the `sub_region_1` level or the `sub_region_2` level - depending on the location you chose, the finer level of granularity may not be available.

```
# TODO Q5
# df_subset =
```

Is the data complete, or is some data not available for the location you have chosen? In the following cell, write code to check for missing data in the `...percent_change_from_baseline` fields.

```
# TODO Q6
# df_subset
```

TODO Q7: Edit this cell to answer the following question: Is the data complete, or is some relevant data missing? Why would some locations only have partial data available (missing some `...percent_change_from_baseline` fields for some dates)? **Include a short quote from the material you read in the "Learn about the data" section to answer this question.**

For this data, the `date` field is important, but we don't necessarily care about the absolute date. Instead, we care about how many days have elapsed since the first confirmed case of COVID-19 in this location, how many days have elapsed since a "stay at home" order or similar rule was established in this location (if there was one) and how many days have elapsed since it was lifted (if applicable).

For example, in Brooklyn, New York, I might compute:

```
days_since_lockdown = (df_subset['date'] - pd.to_datetime('2020-03-20
    00:00:00')).dt.days.values
# NYC lockdown March 20, 2020 https://www.nytimes.com/2020/03/20/us/coronavirus-today.html
```

Compute "days since [some relevant COVID-19 date]" for your location. In a comment, explain the significance of the date you have chosen, and include a link to a news article or other reference supporting the significance of the date. (The news article does not have to be in English.)

```
# TODO Q8
# days_since...
```

Visualize data

Finally, we are going to visualize the changes in human mobility over this time, for the location you have chosen.

In the following cell, create a figure with six subplots, arranged vertically. (You can refer to the example in the "Python + numpy" notebook from this week's lesson.) On the horizontal axis, put the `days_since...` array you computed in the previous cell. On the vertical axes, show:

- `retail_and_recreation_percent_change_from_baseline` in the top subplot
- `grocery_and_pharmacy_percent_change_from_baseline` in the next subplot
- `parks_percent_change_from_baseline` in the next subplot
- `transit_stations_percent_change_from_baseline` in the next subplot
- `workplaces_percent_change_from_baseline` in the next subplot
- `residential_percent_change_from_baseline` in the bottom subplot

```
# TODO Q9
```

TODO Q10: Answer the following question: Do the results seem to satisfy "common sense"? Explain, citing specific data from your plot to support your answer.

TODO Q11: In the Calibrate Region checklist, Google suggests a number of reasons why the data might *not* be useful for understanding the effect of COVID-19-related lockdowns, or why the data might be misleading. For the location you have chosen, briefly answer all of the questions in that checklist. Based on your answers, do you think there are any serious problems associated with using this data for understanding user mobility changes due to COVID-19?