Before you turn in the homework, make sure everything runs as expected. To do so, select **Kernel→Restart & Run All** in the toolbar above. Remember to submit both on **DataHub** and **Gradescope**.

Please fill in your name and include a list of your collaborators below.

In [5]:
```
NAME = "Matthew Brennan"
COLLABORATORS = "Connor McCormick"
```

Project 2: NYC Taxi Rides

Data Ethics

It's important to consider data ethics and the NYC taxi dataset is no exception. In this notebook we will give you some food for thought and prompt you to think critically about some important aspects of this dataset.

Please choose 2 out of the following 4 questions and write thoughtful responses to your chosen questions. You should make sure to backup your responses with proofs, counter examples, real world data, etc. We encourage you to not only draw upon lectures/online resources but also to debate with your peers! Make sure to submit this notebook to datahub.

Questions

1. We downloaded the taxi data from a publicly available endpoint (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)). This means anyone can access this dataset. Can you think of some other external dataset that can be joined with the taxi data that can be utilized to target specific demographics or invade privacy? How could this dataset be used? For example, if we join taxi dataset and street camera, it will be easy to findout who the passenger is.
2. Many people use ride sharing services like Uber or Lyft. Unlike taxi companies, Uber and Lyft know who you are and exactly where you are going. Although the data is not publicy accessible, you can still access part of the data through limited accessed API that's only opened to paid partners (https://developer.uber.com/docs/riders/references/api/v1.2/requests-request_id-get) (https://developer.uber.com/docs/riders/references/api/v1.2/requests-request_id-get)). What can a third party do with these data?

3. If Josh is the chief data scientist at Uber, he has access to all the Uber data. He wants to investigate racial discrimination against drivers by evaluating and performing hypothesis testing on a ride ratings dataset. For example, he can test the hypothesis that given the same level of service, an Asian driver will get different rating from a Caucasian driver. Can Josh perform such experiment? State your assumptions on the data available at Uber.
4. Let's say that the New York City Metropolitan Transportation Authority (NYC MTA) aggregates data from taxis, Ubers, and Lyfts, and determines that people in Brooklyn rarely use taxis/rideshares. The director of the MTA signs a bill to devote all transportation infrastructure funds solely to Manhattan; what is(are) the flaw(s) in making that assumption?

In [4]:
```
response_1 = "Question 1: Through merging our dataset with a Zillow real estate pricing dataset we would
print(response_1)
```

```
Through merging our dataset with a Zillow real estate pricing dataset we would be able to determine an
immense amount of information about each person. The dataset from Zillow would focus upon the differen
t living prices in various regions of Manhattan and we would merge on the latitude and longitudes that
were given in taxi data. By knowing the real estate prices in the area of pickup (or dropoff but we wi
ll focus only on pickup here) we can make a credible guess about the income of the individual. This gu
ess can come in two forms: we can guess based on the notion that the pickup location at certain time,
i.e. 8 am, likely corresponds to home addresses and gauge income accordingly or that we could logicall
y guess income based on areas that one frequently meanders in. However, for the sake of simplicity let
s choose based on homes and time ranges. Then based on an income guess and the overall demographics of
the neighborhood we could create a model about the ethnicity of the individual. Despite not having any
information about ethnicity original, one could postulate about these sorts of details. Information su
ch as this is without a doubt an invasion of privacy because simply the hour of the day and real estat
e information could provide intimate details: income, ethnicity, religion (based on the quarters), and
possibly even family unit. In past endeavors, I have analyzed the distribution of real estate worth in
areas such as the Bay Area, and then first hand observed the areas of focus. From this experiment, I d
eemed that the area of residence can tell a lot about the socioeconomic gap and corresponds very accur
ately to real estate prices in general.
```

In [ ]:
```
# Don't delete me!
```

In [5]:
```
response_2 = "Question 2: A third party that accesses Uber and Lyfts API has unlimited means to gather p
print(response_2)
```

```
Question 2: A third party that accesses Uber and Lyfts API has unlimited means to gather personal info
rmation from this data through the data as a whole but also through merging with various datasets on l
atitude and longitudes of pickup location. Overall, this unethical invasion of privacy poses many conc
erns with one immense issue being that the Uber API documentation discusses the use of General or Priv
eleged scope, and depending on the scope one chooses an individual could have access to a plethora of
information regarding an individual. Some likely information that could be accumulated are about an in
dividuals residence through frequent pickup and dropoff locations at predicted times of the day where
most people are at home, one's income based on the prices of real estate in that area, one's ethnicity
based on the demographics of the area that is deemed as the residence, and further other personal deta
ils. This in turn could spark specefic adds to target distinct interests that pertain to certain ethni
cities (i.e Lulu lemons for a high income area girl). This type of targeting poses many forms of ethic
al concerns and trends toward a racist attitude as far as commercialism, yet the reality are that thes
e forms of adds are an absolute reality when personal information can be bought from companies such as
Uber and Lyft.
```

In [ ]:
```
# Don't delete me!
```

Submission

You're almost done!

Before submitting this assignment, ensure that you have:

1. Restarted the Kernel (in the menubar, select Kernel→Restart & Run All)
2. Validated the notebook by clicking the "Validate" button.

Then,

1. **Submit** the assignment via the Assignments tab in **Datahub**
2. **Upload and tag** the manually reviewed portions of the assignment on **Gradescope**