# Overview of the Assignment:

It is time to start design this is just a draft

**Part 1**: Are you working on your own or with a partner? If with a partner provide their name. If on your own, just state that this is the case.

Individual work.

**Part 2**: Determine the project scope

* In a short paragraph, describe the topic you wish to explore – an update if any
* Update the five business questions that your data warehouse will answer.

The topic is about building a full-stack data warehouse with ETL pipelines, and visualization plots or queries based on the questions we want to answer.

In this project, I plan to use the datasets of NYC Taxi Trip Duration and NYC Airbnb Open Date. These two data are deeply relative to New York City’s tourism activity. The goal is to explore the relationship between taxi passengers, trip duration and Airbnb rentals in NYC and provide insights by using the data warehouse.

Updated business questions:

1. Does people who rent in high average rental rate region also pay high fare for their taxi rides?
2. What is the relation between Airbnb density and the taxi taking rate?
3. How does the average trip duration for taxi rides vary depending on the number of passengers and the pickup/ drop-off location in relation to popular Airbnb rental neighborhoods?
4. What is the most common pickup/ drop-off locations for Airbnb renters?
5. How does the average fare and trip duration for taxi rides vary based on the distance between Airbnb rental locations and popular tourist attractions in the city?

**Part 3:** Data Sources

* Provide two data sources you will be using, for each data source list the number or columns and rows that are in each data source. Provide a header and first 5 rows from each source in separate file.

1. NYC Taxi Trip Duration

Train.csv: (columns: 11, rows: 1458644)

Background pattern

Description automatically generated

Table

Description automatically generated

Test.csv: (columns: 9, rows: 625134)

Table

Description automatically generated

1. New York City Airbnb Open Data

AB\_NYC\_2019.csv (columns: 16, rows: 48895)

A picture containing website

Description automatically generated

Table

Description automatically generated

* What is the URL or location of the data?

1. NYC Taxi Trip Duration: [**https://www.kaggle.com/c/nyc-taxi-trip-duration/data**](https://www.kaggle.com/c/nyc-taxi-trip-duration/data)
2. NYC Airbnb Open Data: [**https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data**](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)

* What information does this data provide that will help answer one or more of the above questions?

The latitude, longitude, and neighborhood in the NYC Airbnb data can let us know the location of the Airbnb. Drop-off longitude, latitude, and drop-off datetime in the NYC Taxi trip Duration can also let us know where the most common destination of the passenger is, and when do they arrive. According to the location, we might know whether there is relation between Airbnb and Taxi taking rate. We might also get to predict the level of convenience in the neighborhood.

* Do you see any issues in the data that will require transformation.

The length (amount) of rows in these two datasets has a huge difference. Therefore, we may need to sample the taxi dataset for better combination. Luckily, both datasets have latitude and longitude columns, which make the analysis more easily.

Also, in taxi dataset, train and test datasets have the different columns, which will need to be integrated.

**Part 4**: Dimensions - Review the data and the business questions from part 2.

* What fields (attributes) are in the data that will be used for the dimensions.

Based on the questions above, we can identify the attributes that can be used for the dimensions from both datasets. In the preliminary stage of selecting attributes, we will focus on

NYC Taxi Trip Duration:

* + pickup\_datetime
  + passenger\_count
  + pickup\_longitude
  + dropoff\_longitude
  + trip\_duration
  + fare\_amount
  + vendor\_id
  + store\_and\_fwd\_flag

NYC Airbnb Open Data:

* + longitude
  + latitude
  + host\_id
  + neighbourhood
  + neighbourhood\_group
  + room\_type
  + price
  + host\_name
* Determine the dimension tables. There should be at least two non-date dimensions and one date dimension for each fact table.

DimLocation:

Location\_id (PK)

Location\_name

Neighborhood

Borough

Latitude

Longitude

Pickup\_location

Dropoff\_location

DimDate:

Date\_id (PK)

Hour

Day

Week

Month

Year

Date

Trip\_duration

DimTaxi:

Taxi\_id (PK)

Vendor\_id

Passenger\_count

Store\_and\_fwd\_flag

Pickup\_location\_key (FK to DimLocation)

Dropoff\_location\_key (FK to DimLocation)

Trip\_duration

Pickup\_location\_key\_prev

Dropoff\_location\_key\_prev

DimAirbnb:

Airbnb\_id (PK)

host\_id

host\_name

Room\_type

Neighborhood

Neighborhood\_group

Latitude

Longitude

Neighborhood\_prev

Neighborhood\_group\_prev

FactTrip:

Trip\_id (PK)

Taxi\_id (FK)

Pickup\_datetime\_key (FK to DimDate)

Dropoff\_datetime\_key (FK to DimDate)

Date\_id (FK)

Distance

Trip\_duration

Fare\_amount

Tip\_amount

Total\_amount

Trip\_distance

Pickup\_location

Dropoff\_location

Location\_id (FK)

Passenger\_count

FactAirbnb:

Id (PK)

Airbnb\_id (FK)

Date\_id (FK)

Availability

Price

Occupancy\_rate

Average\_daily\_rate

Revenue

Available\_rooms

Booked\_rooms

* At least one (non-date) dimension in your design should have a hierarchy.

For the hierarchy, we design a parent-child relationship between boroughs and neighborhoods in DimLocation. The boroughs are the higher-level grouping and neighborhoods are the lower-level grouping.

(e.g., {Boroughs: Manhattan}, {Neighborhoods: Battery Park City})

* What are the attributes that will be tracked via slowly changing dimensions?

For DimTaxi, the borough and neighborhood fields in DimLocation could be tracked using SCD type 2 to capture any changes over time.

For DimAirbnb, the neighborhood field in DimAirbnb could also be tracked using SCD type 2, as neighborhoods can change over time.

* What attributes within the dimensions will need transformation before they are loaded into the dimension, for example it could be to build consistency or any other issues? This is where for example you might build case statements in your code to handle various scenarios. Two to three examples showing some sample data and what you think the transformation will be during your ETL would be helpful here.

1. Date parsing: In the DimDate table, the pickup\_datetime or dropoff\_datetime attribute may need to be parsed into separate columns to aggregate data by different period.
2. Dealing with missing values in some tables can be beneficial to our analysis.

Graphical user interface, table

Description automatically generated Table

Description automatically generated with medium confidence

1. Name standardization: In the DimTaxi table, the vendor’s name attribute may need to be standardized to ensure consistency.
2. Address normalization: In the DimAirbnb table, the address attribute may need to be normalized to make sure the address format is consistency. Also, normalized address can help avoiding duplication. For this problem, I would like to try deriving a new column called pickup\_neighborhood based on the latitude and longitude.
3. In the DimAirbnb table, I found some values in the neighborhood column have the different format.

Table

Description automatically generated

**Part 5**: Facts – Review the data and the business questions from step 1.

* What measurements are in the data that will be used for the fact tables?
* What measures will you be calculating (i.e. using an aggregate function, or some other transformation – recall as an example some of the aggregation you did in assignment 1A)

1. Average fare per pickup location

Average rental rate per pickup location

Correlation coefficient between average rental rate and average fare per pickup location

Code example:

SELECT

DimTaxi.vendor\_id,

AVG(FactTaxi.fare\_amount) AS avg\_fare,

AVG(FactTaxi.trip\_duration) AS avg\_duration

FROM

DimTaxi

INNER JOIN FactTaxi ON DimTaxi.taxi\_id = FactTaxi.taxi\_id

GROUP BY

DimTaxi.vendor\_id

1. Total number of Airbnb rentals per pickup location

Total number of taxi rides per pickup location

Code example:

SELECT

DimAirbnb.neighborhood,

COUNT(FactAirbnb.listing\_id) AS total\_listings,

AVG(FactAirbnb.price) AS avg\_rental\_rate

FROM

DimAirbnb

INNER JOIN FactAirbnb ON DimAirbnb.listing\_id = FactAirbnb.listing\_id

GROUP BY

DimAirbnb.neighborhood

1. Average trip duration per number of passengers

Average trip duration per pickup location and Airbnb rental neighborhood

1. Percentage of Airbnb rentals that occur at these pickup/ drop-off locations.
2. Average fare per distance between Airbnb location and popular tourist spots.

Average trip duration per distance between Airbnb location and popular tourist spots.

**Part 6**: Design – Create a Draw.io, Visio or Lucidchart diagram of your constellation data warehouse design.

Diagram, schematic

Description automatically generated

Use the **Ask the Teaching Team Discussion Forum** if you have any questions regarding the how to approach this assignment.

Save your assignment as ***lastnameFirstname\_ProjectUpdate2.docx*** and submit it in the *Assignments* section of the course.

For help uploading files please refer to the *Technical Support* page in the syllabus.

Project scoping is graded based on the following:

1 – On track, 0-Off track, .5 – partially on track