# Overview of the Assignment:

Let’s focus on the ETL Process.

Fist iteration of ETL: Take a part of your design and implement ETL. The scope at this point is limited and should include just two of your dimensions and one fact table. Using Python, SQL document and walk through the initial process of loading your data into staging (SQL staging or a data frame) doing some transformation and loading into your dimensions and a fact table.

**Part 7**: Provide code and screenshots of loading your data into staging/data frame\

*#!/usr/bin/env python*

*# coding: utf-8*

*# In[32]:*

**import** pandas **as** pd

**import** numpy **as** np

**import** pyodbc **as** odbc

*# In[69]:*

Airbnb\_df **=** pd.read\_csv('AB\_NYC\_2019.csv')

Airbnb\_df.head()

*# In[3]:*

Taxi\_df **=** pd.read\_csv('train.csv')

Taxi\_df.head()

*# In[4]:*

Taxi\_df2 **=** pd.read\_csv('test.csv')

Taxi\_df2.head()

*# In[6]:*

*# Combine the train and test df as Taxi\_df*

*# Add a new column to identify the source*

Taxi\_df['source'] **=** 'train'

Taxi\_df2['source'] **=** 'test'

*# Reset Index*

Taxi\_df2 **=** Taxi\_df2.reindex(columns**=**Taxi\_df.columns)

*# Concatenate train and test dataframe*

Taxi\_df **=** pd.concat([Taxi\_df, Taxi\_df2], ignore\_index**=True**)

*# Fill any missing value with NaN*

Taxi\_df['trip\_duration'] **=** Taxi\_df['trip\_duration'].fillna(**-**1)

*# Save the new df as CSV file*

Taxi\_df.to\_csv('NYC\_Taxi.csv', index**=False**)

*# In[7]:*

*# Check the new df*

Taxi\_df.head()

*# In[8]:*

*# Total rows in our target dfs*

Taxi\_total\_rows **=** Taxi\_df.shape[0]

Airbnb\_total\_rows **=** Airbnb\_df.shape[0]

print('Taxi total rows: ', Taxi\_total\_rows, '\nAirbnb total rows: ', Airbnb\_total\_rows)

*# In[11]:*

*# Count the total columns*

Taxi\_total\_cols **=** Taxi\_df.shape[1]

Airbnb\_total\_cols **=** Airbnb\_df.shape[1]

print('Taxi total columns: ', Taxi\_total\_cols, '\nAirbnb total columns: ', Airbnb\_total\_cols)

*# In[14]:*

*# Dtypes*

Airbnb\_df.dtypes

*# In[15]:*

Taxi\_df.dtypes

*# In[19]:*

*# PART 2 - Creating SCD Type 2 and 3 Dim and Key Maintenance*

*# dfs Cleaning*

Airbnb\_df.isnull().sum()

*# Duplication check*

duplicate\_counts **=** Airbnb\_df.groupby(['id','host\_id','host\_name']).size().reset\_index(name**=**'Count')

*# Filter the rows where the count is greater than 1 to find the combinations*

Airbnb\_duplicate **=** duplicate\_counts[duplicate\_counts['Count'] **>** 1]

print("Duplicate combinations of Columns in Airbnb:\n")

print(Airbnb\_duplicate['Count'].sum())

print('\n',Airbnb\_duplicate)

*# In[43]:*

*# Data normalization*

Airbnb\_df['host\_name'] **=** Airbnb\_df['host\_name'].str.slice(stop**=**255)

Airbnb\_df['host\_name'] **=** Airbnb\_df['host\_name'].str.encode('ascii', errors**=**'replace').str.decode('ascii')

Airbnb\_df['name'] **=** Airbnb\_df['name'].str.slice(stop**=**255)

Airbnb\_df['name'] **=** Airbnb\_df['name'].str.encode('ascii', errors**=**'replace').str.decode('ascii')

Airbnb\_df['last\_review'] **=** pd.to\_datetime(Airbnb\_df['last\_review'])

Airbnb\_df['reviews\_per\_month'] **=** Airbnb\_df['reviews\_per\_month'].fillna(0)

*# In[45]:*

*# Filter and load data*

Airbnb\_filtered **=** Airbnb\_df[['id','name','host\_id','host\_name','neighbourhood\_group','neighbourhood',

                            'room\_type','longitude','latitude','price','calculated\_host\_listings\_count',

                             'availability\_365','last\_review','reviews\_per\_month']]

Airbnb\_filtered.head()

*#Airbnb\_filtered.dtypes*

*# In[66]:*

*# Estiblish the connection*

conn **=** odbc.connect(Trusted\_Connection **=** 'YES',

                    Driver **=** '{ODBC Driver 17 for SQL Server}',

                    Server **=** 'DESKTOP-IAN\JLM\_SQLSERVER',

                    Database **=** 'CS689\_fINALpROJ')

print(conn)

*# In[67]:*

*# Create table*

cursor **=** conn.cursor()

cursor.execute("""

CREATE TABLE Airbnb\_staging (

    airbnb\_id INT PRIMARY KEY,

    name NVARCHAR(255),

    host\_id INT,

    host\_name NVARCHAR(255),

    neighborhood\_group NVARCHAR(255),

    neighborhood NVARCHAR(255),

    room\_type NVARCHAR(255),

    longitude FLOAT,

    latitude FLOAT,

    price INT,

    calculated\_host\_listings\_count INT,

    availability\_365 INT,

    last\_review DATETIME,

    reviews\_per\_month FLOAT

)

""")

conn.commit()

conn.close()

*# In[68]:*

*# Load data*

conn **=** odbc.connect(Trusted\_Connection **=** 'YES',

                    Driver **=** '{ODBC Driver 17 for SQL Server}',

                    Server **=** 'DESKTOP-IAN\JLM\_SQLSERVER',

                    Database **=** 'CS689\_fINALpROJ')

cursor **=** conn.cursor()

**for** row **in** Airbnb\_filtered.itertuples():

**try**:

        cursor.execute("""

            INSERT INTO Airbnb\_staging (

                airbnb\_id,

                name,

                host\_id,

                host\_name,

                neighborhood\_group,

                neighborhood,

                room\_type,

                longitude,

                latitude,

                price,

                calculated\_host\_listings\_count,

                availability\_365,

                last\_review,

                reviews\_per\_month

            )

            VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)

            """,

            row.id,

            row.name **if** **not** pd.isna(row.name) **else** 'Unknown',

            row.host\_id,

            row.host\_name **if** **not** pd.isna(row.host\_name) **else** 'Unknown',

            row.neighbourhood\_group,

            row.neighbourhood,

            row.room\_type,

            float(row.longitude),

            float(row.latitude),

            row.price,

            row.calculated\_host\_listings\_count,

            row.availability\_365,

            row.last\_review **if** **not** pd.isna(row.last\_review) **else** **None**,

            row.reviews\_per\_month **if** **not** pd.isna(row.reviews\_per\_month) **else** **None**

        )

**except** odbc.ProgrammingError **as** e:

        print(**f**"Error occurred for row: {row}")

        print(**f**"Error message: {e}")

**break**

conn.commit()

conn.close()

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

SELECT \* FROM Airbnb\_staging;

A screenshot of a computer

Description automatically generated

**Part 8:** Provide code and screenshots of transforming the data. Perhaps you are adjusting for consistency of data or calculating aggregates.

In the Python part, we already have filtered some of the data in Airbnb\_df, such as NA value check and columns’ normalization, such as transform the data type of last\_review into dattime format and so on.

A screenshot of a computer code

Description automatically generated with low confidence

A screen shot of a computer program

Description automatically generated with medium confidence

DELETE FROM Airbnb\_staging

WHERE last\_review IS NULL;

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Since last\_review will become an important parameter for us to create the measures between DimDate and FactAirbnb, so we dropped all the NULL values in last\_review.

Before dropped: 48895 rows.

After dropped: 38843 rows.

**Part 9:** Provide code and screenshots of loading the data into the two dimensions and the fact. At this time, you do not need to worry about maintenance of slowly changing dimensions, the focus is on the initial data load. If you are loading into SCD2 or SCD3, make sure to show the SCD maintenance attributes populated.

-- Create DimAirbnb\_Property

CREATE TABLE DimAirbnb\_Property (

Airbnb\_id INT IDENTITY(1,1) PRIMARY KEY,

Leasing\_id INT,

Room\_name NVARCHAR(255),

Room\_type NVARCHAR(50),

Price INT,

Neighborhood NVARCHAR(255),

Neighborhood\_group NVARCHAR(255),

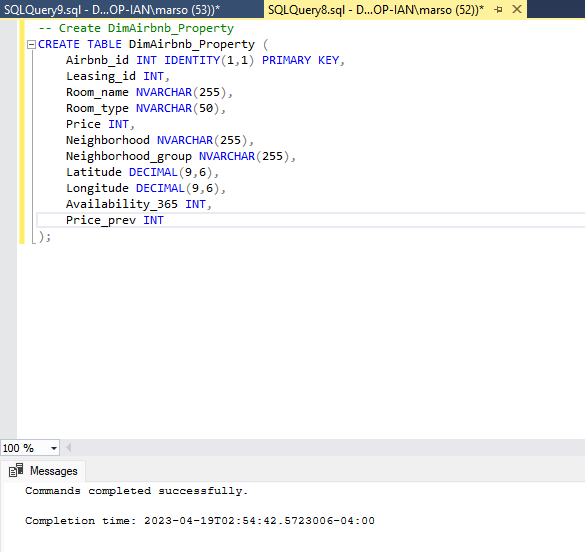
Latitude DECIMAL(9,6),

Longitude DECIMAL(9,6),

Availability\_365 INT,

Price\_prev INT

);



-- Insert data into DimAirbnb\_Property

INSERT INTO DimAirbnb\_Property (Airbnb\_id, Room\_name, Room\_type, Price, Neighborhood, Neighborhood\_group, Latitude, Longitude, Availability\_365)

SELECT airbnb\_id, name, room\_type, price, neighborhood, neighborhood\_group, latitude, longitude, availability\_365

FROM Airbnb\_staging;

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

-- Create DimAirbnb\_Host

CREATE TABLE DimAirbnb\_Host (

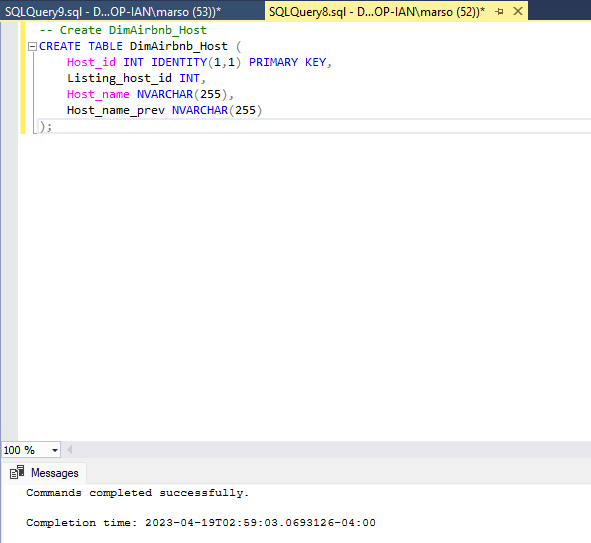
Host\_id INT IDENTITY(1,1) PRIMARY KEY,

Listing\_host\_id INT,

Host\_name NVARCHAR(255),

Host\_name\_prev NVARCHAR(255)

);



-- Insert data into DimAirbnb\_Host

INSERT INTO DimAirbnb\_Host (Listing\_host\_id, Host\_name)

SELECT DISTINCT host\_id, host\_name

FROM Airbnb\_staging;

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

CREATE TABLE DimDate (

Date\_id INT IDENTITY(1,1) PRIMARY KEY,

DateHour DATETIME NOT NULL,

Hour INT NOT NULL,

Day INT NOT NULL,

Month INT NOT NULL,

Year INT NOT NULL,

Week INT NOT NULL

);

A screenshot of a computer

Description automatically generated with medium confidence

WITH DateRange (DateHour) AS (

SELECT CAST('2010-01-01 00:00:00' AS DATETIME)

UNION ALL

SELECT DATEADD(hour, 1, DateHour)

FROM DateRange

WHERE DateHour < '2023-12-31 23:00:00'

)

INSERT INTO DimDate (DateHour, Hour, Day, Month, Year, Week)

SELECT

DateHour,

DATEPART(hour, DateHour) AS Hour,

DAY(DateHour) AS Day,

MONTH(DateHour) AS Month,

YEAR(DateHour) AS Year,

DATEPART(wk, DateHour) AS Week

FROM DateRange

OPTION (MAXRECURSION 0);

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

-- Create FactAirbnb

CREATE TABLE FactAirbnb (

Id INT IDENTITY(1,1) PRIMARY KEY,

Airbnb\_id INT FOREIGN KEY REFERENCES DimAirbnb\_Property(Airbnb\_id),

Date\_id INT FOREIGN KEY REFERENCES DimDate(Date\_id),

Availability INT,

Price INT,

Revenue DECIMAL(10,2),

Available\_rooms INT,

Booked\_rooms INT,

Occupancy\_rate DECIMAL(5,2),

Average\_daily\_rate DECIMAL(10,2)

);

A screenshot of a computer

Description automatically generated with medium confidence

-- Insert data into FactAirbnb

INSERT INTO FactAirbnb (Airbnb\_id, Date\_id, Price)

SELECT

DP.Airbnb\_id,

DD.Date\_id,

ASG.price

FROM

Airbnb\_staging ASG

INNER JOIN DimDate DD ON ASG.last\_review = DD.DateHour

JOIN DimAirbnb\_Property DP ON ASG.Airbnb\_id = DP.Leasing\_id;

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

(The measures in FactAirbnb are still developing by utilize the availability\_365 column)

If you update any of the previous parts and would like feedback for a specific part, please make a note:

Updates in PART 4, PART 5, and PART 6.

**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
  + calculated\_host\_listings\_count
  + name
  + last\_review
  + review\_per\_month
  + availability\_365
* Determine the dimension tables. There should be at least two non-date dimensions and one date dimension for each fact table.

Dimension table:

DimDate:

Date\_id (PK)

Hour

Day

Week

Month

Year

DateHour

DimLocation:

Location\_id (PK)

Location\_type

Neighborhood

Borough

Latitude

Longitude

DimTaxi:

Taxi\_id (PK)

Vendor\_id

Passenger\_count

Pickup\_datetime

Dropoff\_datetime

Store\_and\_fwd\_flag

Trip\_duration

Effective\_date

End\_date

Is\_current

DimAirbnb\_Property:

Airbnb\_id (PK)

Leasing\_id (original Airbnb\_id)

Room\_name

Room\_type

Price

Neighborhood

Neighborhood\_group

Latitude

Longitude

Availability\_365

Price\_prev

DimAirbnb\_Host:

Host\_id (PK)

Listing\_host\_id (original Host\_id)

Host\_name

Host\_name\_prev

Fact table:

FactTrip (Cumulative)

Trip\_id (PK)

Taxi\_id (FK)

Location\_id (FK)

Date\_id (FK)

Pickup\_datetime\_key (FK to DimDate)

Dropoff\_datetime\_key (FK to DimDate)

Pickup\_location (FK to DimLocation)

Dropoff\_location (FK to DimLocation)

Passenger\_count

Trip\_distance

Trip\_duration

Fare\_amount

Total\_amount

FactAirbnb (Snapshot)

Id (PK)

Airbnb\_id (FK)

Date\_id (FK)

Host\_id (FK)

Availability

Price

Revenue

Available\_rooms

Booked\_rooms

Occupancy\_rate

Average\_daily\_rate

FactTaxiMonthlySummary (Cumulative)

Sum\_id (PK)

Month\_id (FK to DimDate)

Host\_id (FK)

Vendor\_id

Total\_passengers

Total\_trip\_duration

Total\_revenue

* 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})

Also, in DimDate, we have a hierarchy of time from Year, month, day to time.

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

For DimAirbnb\_Property table, the price could be tracked by using SCD Type 3 with new column for the price to capture any changes over time.

Also, in the DimAirbnb\_Host table, we can use SCD Type 3 to capture the host name changes.

For DimTaxi table, we added effective date and end date as SCD Type 2 to record any changes in vendor id.

DimLocation using SCD Type 1 would be enough.

* 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.
4. 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.
5. 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?

Trip\_duration

Trip\_distance

Total\_passengers

Total\_trip\_duration

Total\_revenue

Available\_rooms

Booked\_rooms

Average\_daily\_rate

Occupancy\_rate

* 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

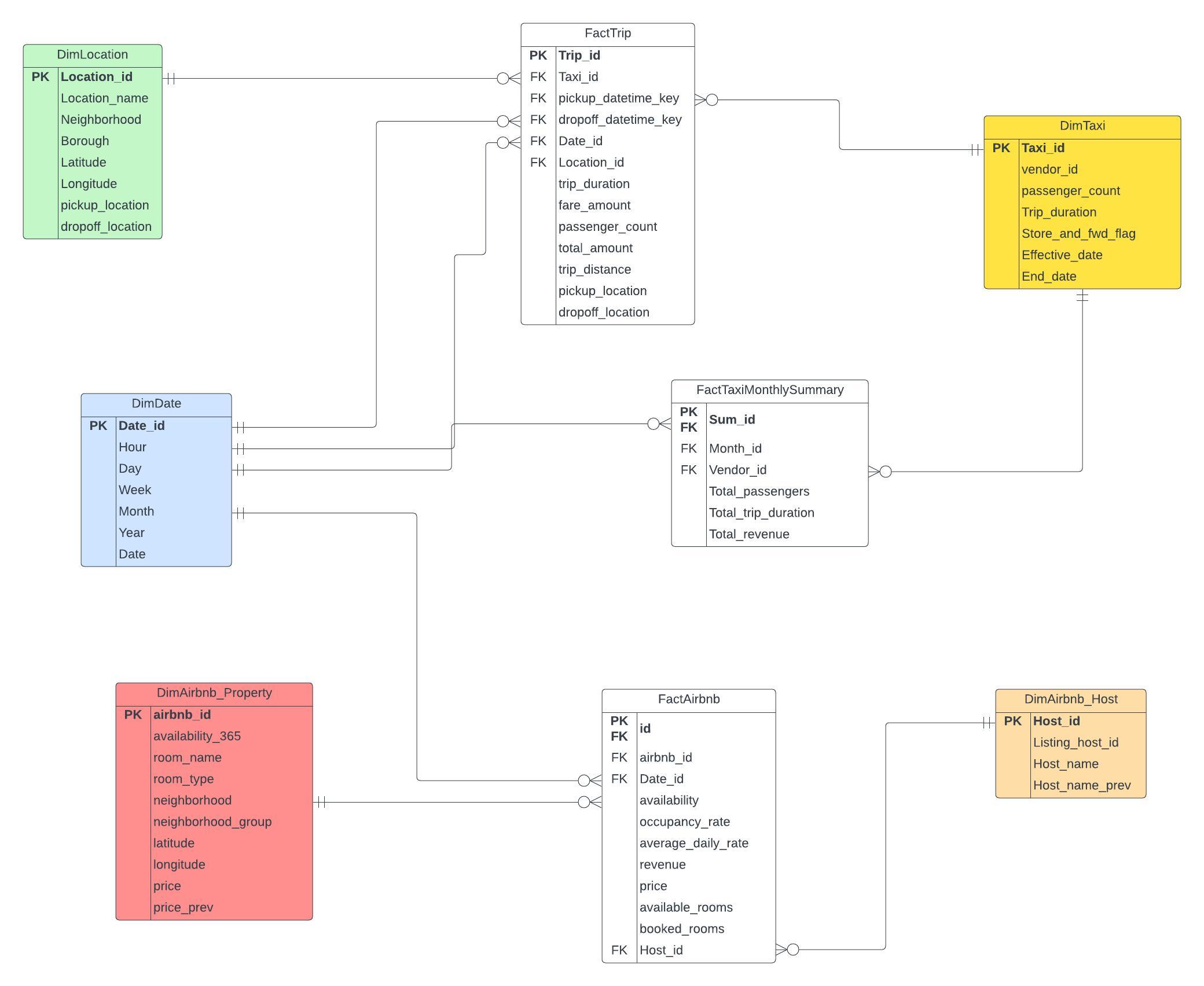
1. Average trip duration per number of passengers

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.



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