Team 3_Team Project_Phase II: **Airbnb Booking Demand**

1. **State the problem**
   Short-term rentals are an important part of Boston's housing and tourism market, and Airbnb hosts care a lot about how often their listings get booked. In this project, we study how well we can predict the occupancy rate of individual Airbnb listings in Boston using datasets from InsideAirbnb the calendar, listings, and reviews data.
   - Target Variable: Occupancy Rate
     - Define occupancy over the period 2016-09-05 to 2017-09-04 as:

       $$\textbf{occupancy} = 1 - \textbf{AVG(availability)}$$

     - In the calender data:
       - 't' = 1 means the listing is available
       - 'f' = 0 means the listing is booked

   - Prediction target is this listing-level occupancy rate between 0 and 1.
   - Unit: Listing
     - Calendar and review data are first summarized at the listing level and then merged with listing attributes.

2. **Tell us who cares about this problem and Why**
   Predicting occupancy is practically essential for several groups:
   - **Hosts and property managers.**
     They rely on expected occupancy to decide nightly prices, minimum-night rules, and renovation priorities. Even a rough forecast can help reduce empty nights and smooth income.
   - **Platform operators such as Airbnb**.
     The platform can use occupancy models to design search ranking, dynamic pricing tools, and host onboarding strategies by neighborhood and season.
   - **Short-term rental investors and real-estate developers.**
     Occupancy expectations enter directly into revenue projections when evaluating whether a new listing or building is financially viable.
   - **City planners and tourism bodies.**
     Understanding which neighborhoods sustain high occupancy over time helps them monitor housing pressure and plan infrastructure for tourism.

   Because occupancy is hard to predict in advance and is closely tied to revenue and regulation, building and evaluating prediction models on real Boston data is useful both academically and in practice.

3. **Describe your data – where it came from, what it contains**

   We will use the Boston Airbnb Open Data from Kaggle. This dataset is part of InsideAirbnb, and the source can be found at Get the Data | Inside Airbnb. The data we are using range from 2016-09-05 to 2017-09-04. They were collected because guests and hosts have used Airbnb to travel in a more unique, personalized way since 2008. As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in Boston, MA.

- **Calendar.csv**: 1,048,576 rows, 4 columns, type: int, object, datetime
  - We use this file to compute occupancy and to summarize prices over the one-year period.
- **Listing.csv**: 3585 rows, 95 columns, type: int, float, object
  - In our models, we mainly use: id, price, room_type, property_type, accommodates, bathrooms, bedrooms, beds, number_of_reviews, review_scores_rating, reviews_per_month, host_is_superhost, host_total_listings_count, neighbourhood_cleansed, latitude, longitude.
  - These variables describe the physical characteristics of the listing, host experience, reputation, and location.
- **Reviews.csv**: 68,275 rows, 6 columns, type: int, object
  - Contains listing_id, review date, and text comments.

We restrict all three files to the same one-year window to keep the target and features aligned in time.

4. Present some interesting descriptive analyses (plots/tables) that motivate your exercise
5. Present your main results
6. Which methods worked best for your problem?
7. What were the challenges you faced? Tell us about the biggest challenge you faced and how you overcame it (or, tried but did not – that's fine too – not every problem has a solution.)
8. Conclude – what have you learnt that can be put to practice?

These points each have approximately equal weight in your final score.