# Dataset Collection Documentation

## Introduction

Two datasets, “UnitToBasePrice” and “Bookings” are collected in order to supplement ML model training to predict room price. The first dataset provides an average price per day of a BukitVista *unit* over many different bookings, accompanied with features corresponding to that *unit*. The second dataset provides the final sold price per day of a *unit* for each booking, while accompanied by features corresponding to that booking and the avaergae price per day obtained from the previous dataset.

# Source

Dataset obtain from <https://datavista.bukitvista.com>, access given by BukitVista.

Link to first dataset’s query: : [Github link]

Link to second dataset’s query: : [Github link]

# Data Collection Methodology

## UnitToBasePrice

1. From *booking* table in datavista, obtain booking’s id, check in, check out, and date for when the booking was created, it’s earnings and currency, as well as unit id and room id obtained as a LEFT JOIN from another table *booking\_to\_units* in order to attach a booking to a certain unit.
2. Convert all earnings value to IDR utilizing the *exchange* table, a time series table containing daily exchange rates.
3. Calculate earnings per day as a new column by dividing converted earnings by the number of days between check out and checkin.
4. Calculate the average of per day price for a single unit, over all bookings.
5. Attach existing columns with features corresponding to the unit from tables *unit* (number of bedroom, bathroom, beds and capacity) *unit\_amenities* (various amenities in boolean values) and *unit*

## Bookings

1. Follow UnitToBasePrice steps 1 to 3
2. From check in and check out values, calculate length of stay by subtracting as a new column.
3. From created\_at and check\_in, calculate booking window by subtracting as a new column.
4. Obtain columns from *review* table and left join with it’s corresponding booking id.

# Data Cleaning and Preprocessing

# Data Storage

Format : Comma Seperated Values

Stored in Github.