1. **Structure**
   1. EDA *(Minji)*
   2. CV LASSO / Ridge (mainly numerical & categorical?) *(Sarah)*
   3. Text Analysis (variables: {description, name}) *(Elena)*
   4. PCA first, then K-means Cluster *(Yuting*)
   5. KNN? / Random Forest *(Mengdi)*
2. **Data Preprocessing**
   1. **Type of variables**

* **Numerical:**

log\_price;

accommodates;

bathrooms (200, 0.27% missing);

host\_response\_rate (18299, 24.7% missing);

latitude;

longitude;

number\_of\_reviews;

review\_scores\_rating (16722, 22.6% missing);

bedrooms (91, 0.12% missing);

beds (131, 0.18% missing)

* **Categorical:**

property\_type (20+ categories);

room\_type (3 categories);

amenities (dictionary); ***🡪 turn into multiple binary variable columns***

bed\_type (5 categories);

cancellation\_policy (5 categories);

cleaning\_fee (True and False);

city (6 categories);

host\_has\_profile\_pic (True or False) (188, 0.25% missing);

host\_identity\_verified (True or False) (188, 0.25% missing);

instant\_bookable (True or False);

neighbourhood (200+ categories) (6872, 9.27% missing);

zipcode (200+ categories) (966, 1.3% missing)

* **Textual:**

description; ***🡪 only used in text analysis***

name; ***🡪 only used in text analysis***

* **Date:**

first\_review (15864, 21.4% missing); ***🡪 turn into days\_since\_first\_review***

last\_review (15827, 21.4% missing); ***🡪 turn into days\_since\_last\_review***

host\_since (188, 0.25% missing); ***🡪 turn into host\_duration***

* **URL:**

thumbnail\_url (8216, 11.1% missing)

***🡪 turn into a binary variable indicating whether it has an URL or not***