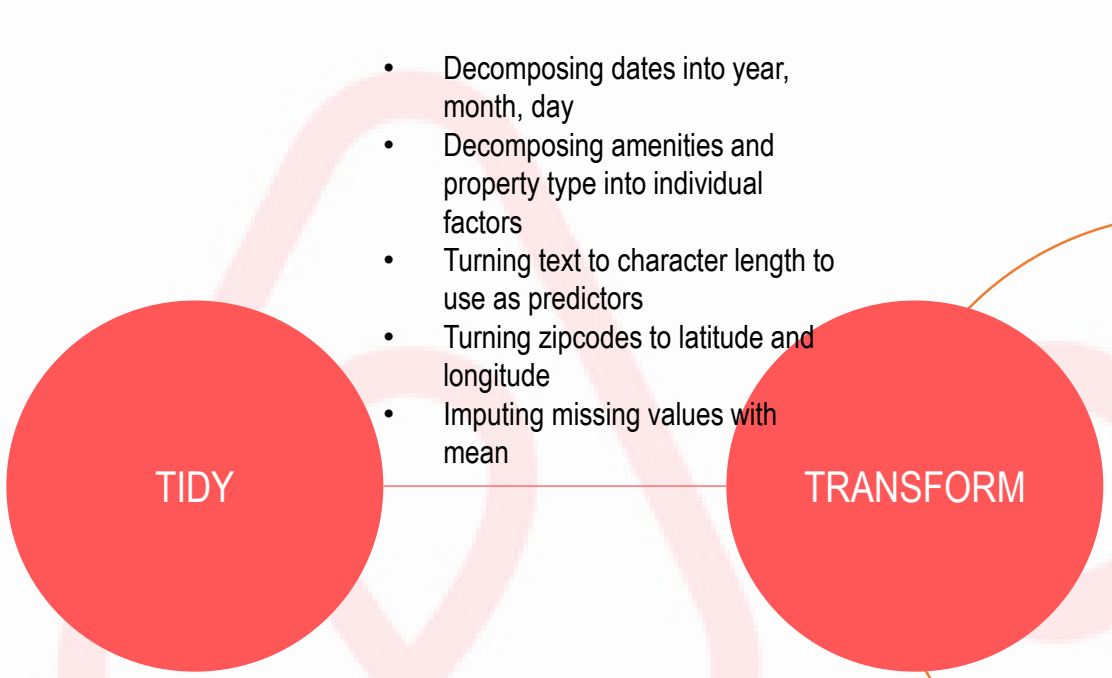
Predicting Airbnb Prices

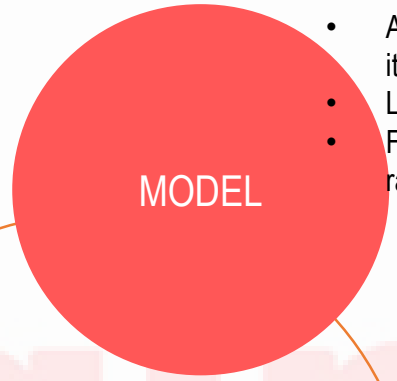
Charlie Mei
(cm3947)

MODELING FRAMEWORK

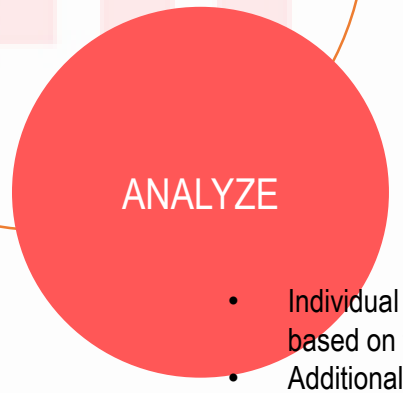


- Split to training and test data
- Organize predictors into categories – about the host, location, about the property

- Decomposing dates into year, month, day
- Decomposing amenities and property type into individual factors
- Turning text to character length to use as predictors
- Turning zipcodes to latitude and longitude
- Imputing missing values with mean



- Predictors added iteratively by categories
- Additional predictors added iteratively to improve RMSE
- Linear regression used first
- Final technique used: random forest



- Individual models were compared based on test RMSE
- Additional variables added depending on whether they fitted theory

FINAL MODEL

- Predictors: 69
- Predictors considered per tree: 30
- Three-fold cross validation
- 500 trees

TAKEAWAYS

- Stick to a modeling framework to guide the modeling process
- Using separate cleaning and modeling script files

COMPLICATIONS

- Linear regression models cap out up to a certain point
- Removing price outliers significantly increased RMSE
- Random forest models are computationally expensive to run

NEXT TIME

- Deeper text analysis to extract meaningful predictors
- Taking a deeper look at outliers and understanding whether they should be removed
- Considering using other machine learning models