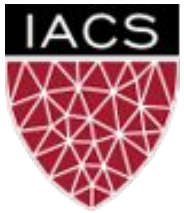
REX House Discovery

Milestone 1

Weiru Chen, Nikhil Vanderklaauw,
Zhenwei (Selina) Wu, Wanxi (Cecilia) Yang



INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

REX
REAL ESTATE

10/04/2020

Problem Statement

REX wants to improve users' experience in buying and homes.

Goal: Develop an application that serves open minded house-hunters with personalized matches for discovering their perfect home

Scope of Work

- **Focusing on**
  - Generate a **user taste profile**
  - **Utilize** and improve upon **REX's** pre-existing **tagging algorithms** (NLP , CV)
  - Build **custom recommendation system** (CF), stretch goal, Dynamic - updates based on user-feedback
- **Not focusing on**
  - Application building
  - UI design
  - New user experience study

Note: In Milestone 1, all EDA and modeling are based on general structured data.

Team

	Zhenwei	Nikhil	Weiru	Wanxi
EDA on csv files	House market list	Assessor	Transaction	Join house market list
	Data analysis		Baseline model	

Learning goals

Techniques

- Do exploratory data analysis on real-world datasets
- Learn about recommendation models
  - Content-based filtering
  - Collaborative filtering
- Integrate models into working product

Others

- Understand the most important aspects in home-selecting and real estate transactions
- Communicate with the client and understand their needs to improve their product/business

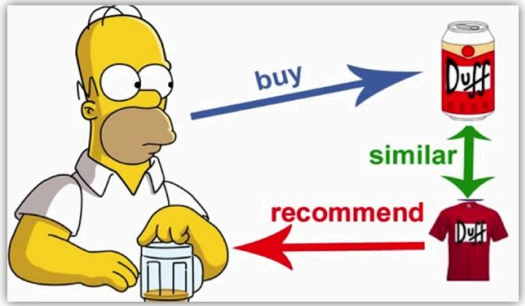
Relevant Knowledge

- Collaborative Filtering
  - Build a Recommendation Engine With Collaborative Filtering
  - (https://realpython.com/build-recommendation-engine-collaborative-filtering/)

	i_1	i_2	i_3	i_4	i_5
u_1	5		4	1	
u_2		3		3	
u_3		2	4	4	1
u_4	4	4	5		
u_5	2	4		5	2

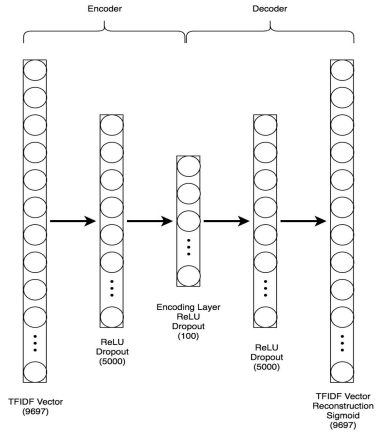
Rating Matrix

- Content-based Filtering
  - Recommendation Systems with Python: Content-Based Filtering
  - (https://heartbeat.fritz.ai/recommender-systems-with-python-part-i-content-based-filtering-5df4940bd831)

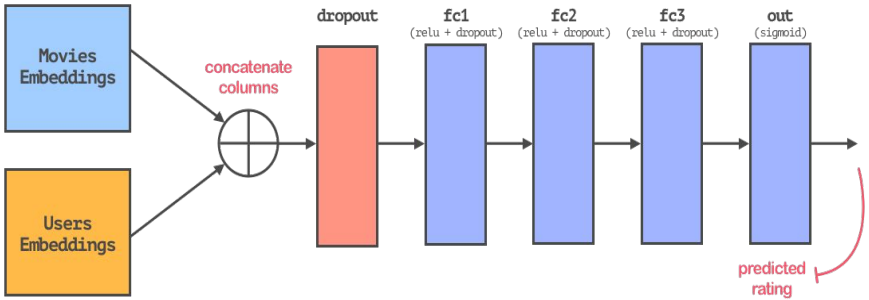


Relevant Knowledge

- Content-based Filtering with ML/DL models
  - Creating a Hybrid Content-Collaborative Movie Recommender Using Deep Learning
  - (https://towardsdatascience.com/creating-a-hybrid-content-collaborative-movie-recommender-using-deep-learning-cc8b431618af)



- Collaborative Filtering with ML/DL models
  - How to Implement a Recommendation System with Deep Learning and PyTorch
  - (https://medium.com/coinmonks/how-to-implement-a-recommendation-system-with-deep-learning-and-pytorch-2d40476590f9)



- Hybrid of Collaborative & Content-based Filtering with ML/DL models

Project Ideas

Step 1: Explore the datasets we have

Step 2: Create a model to link similar homes with key features

Step 3: Improve the model with real user data, NLP, and image processing

Step 4: Integrate the model

Data Available

Focus on Suffolk County (FIPS 25025) - Boston, Allston, Brighton, JP etc.

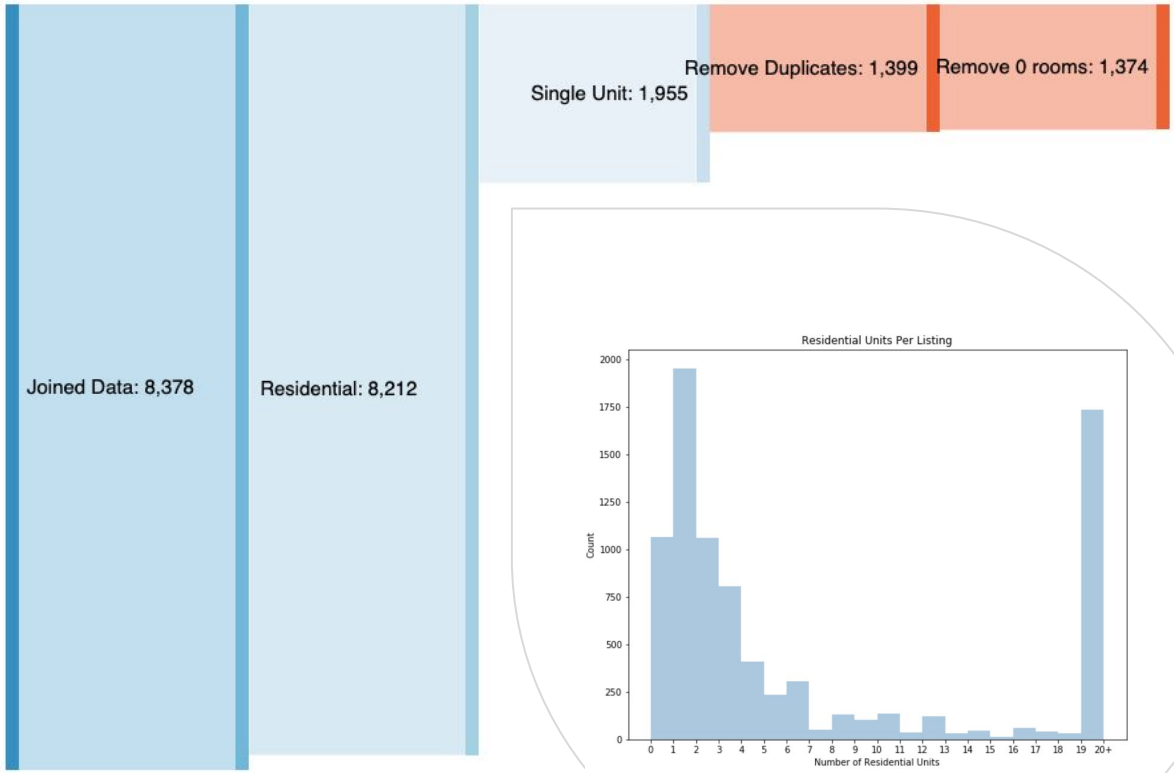
- First American Assessor data (162 features, 20k items)
  - Census of raw building data
  - Standardized across counties

- House_market_listing_property (83 features, 42k items)
  - Multiple Listing Service (MLS) + proprietary

Exploratory Data Analysis

Data Filtering Funnel

1. Property Type
2. Multiple units
3. Duplicate addresses
4. 0 Rooms



How to treat missing values?

We **drop** columns with more than **30% missing** values. Remove all rows with **no bedroom and bathroom** information

For some variables, could replace “missing” with proxy columns

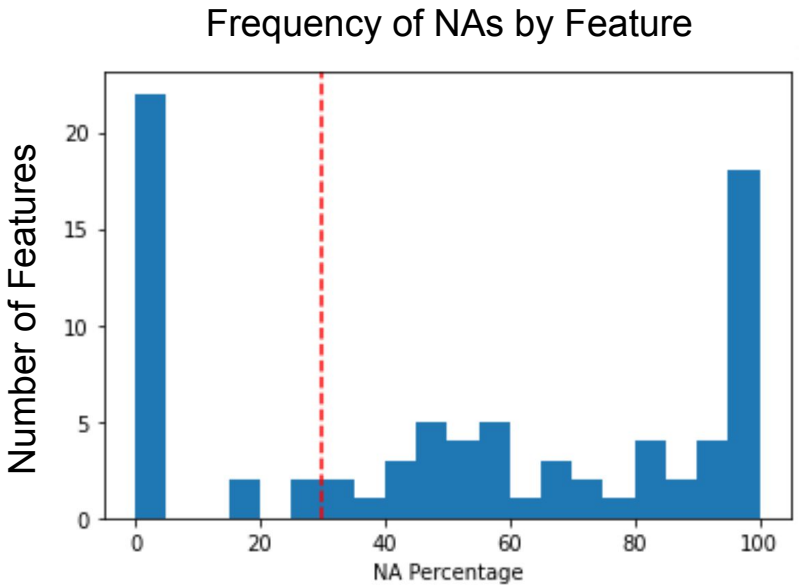
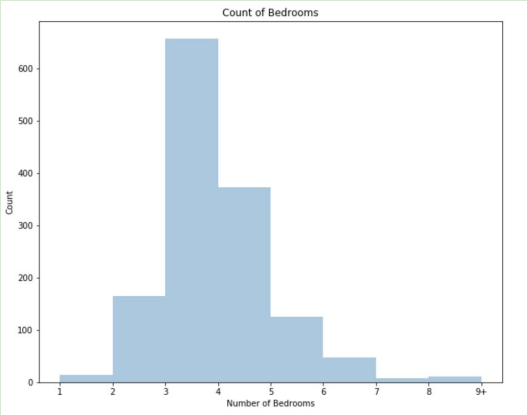


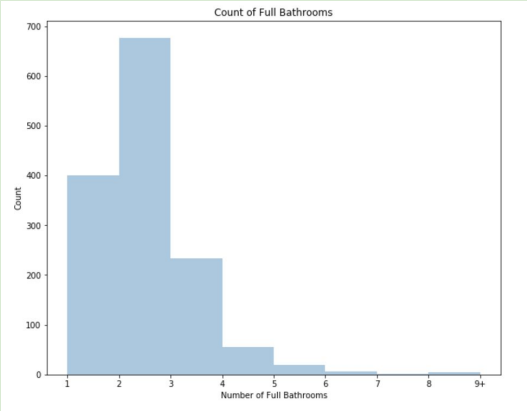
Fig. Missing value histogram for table: house_market_listing_property

Exploratory Data Analysis: What kind of houses do we have in Suffolk County?

Home Stats

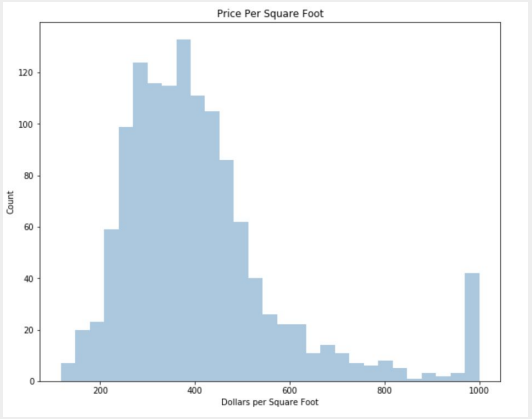


Bedrooms

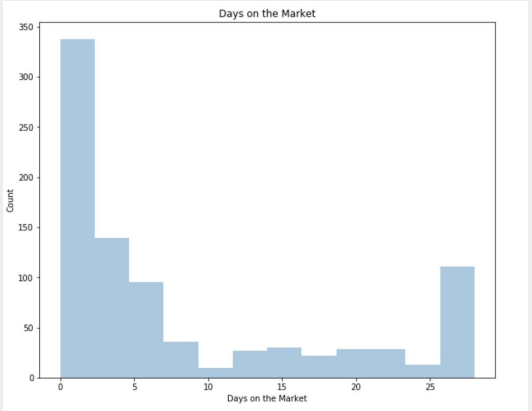


Bathrooms

The Market



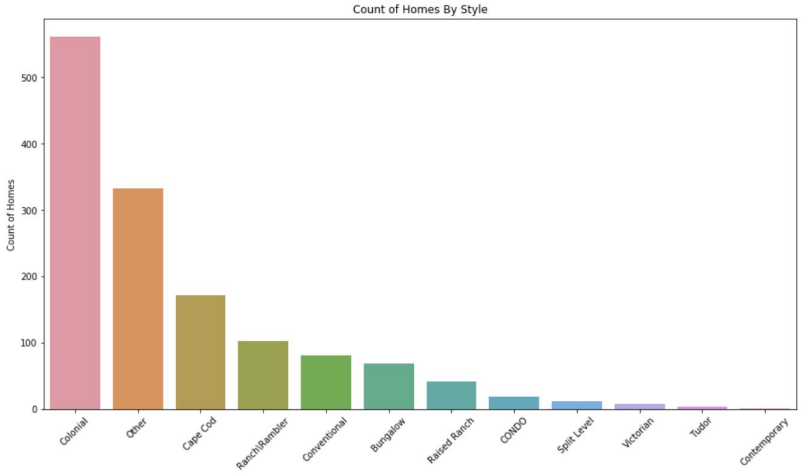
Price Per SqFt



Days on Market

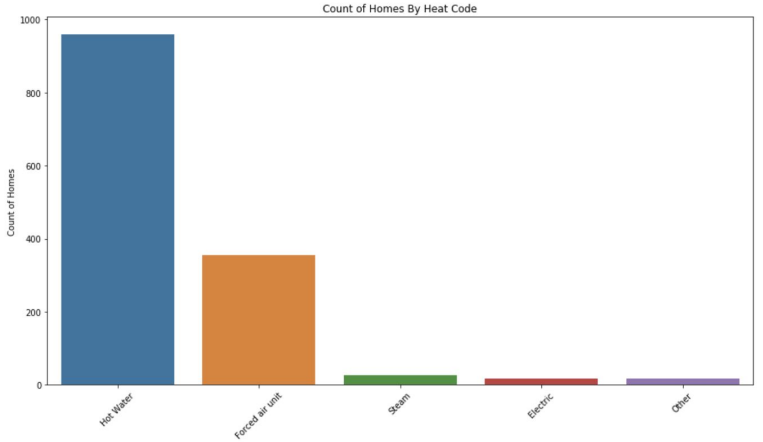
Exploratory Data Analysis: What kind of houses do we have in Suffolk County?

Home Features

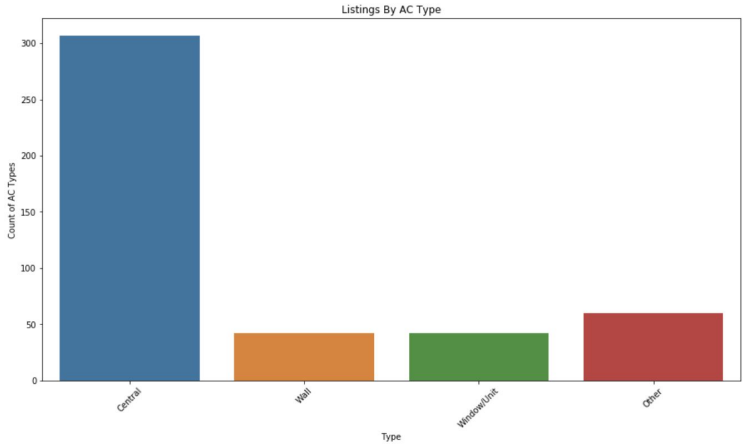


Home Style

Heating Type



AC Type



Other sought after amenities:

- 33 have pools
- 99 have garages

What kind of houses do we have in Suffolk County?

Year Built & Bedroom Count

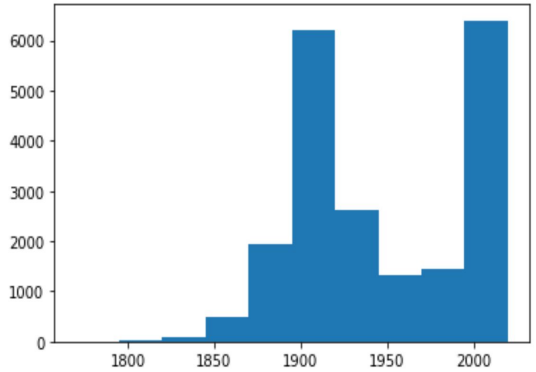
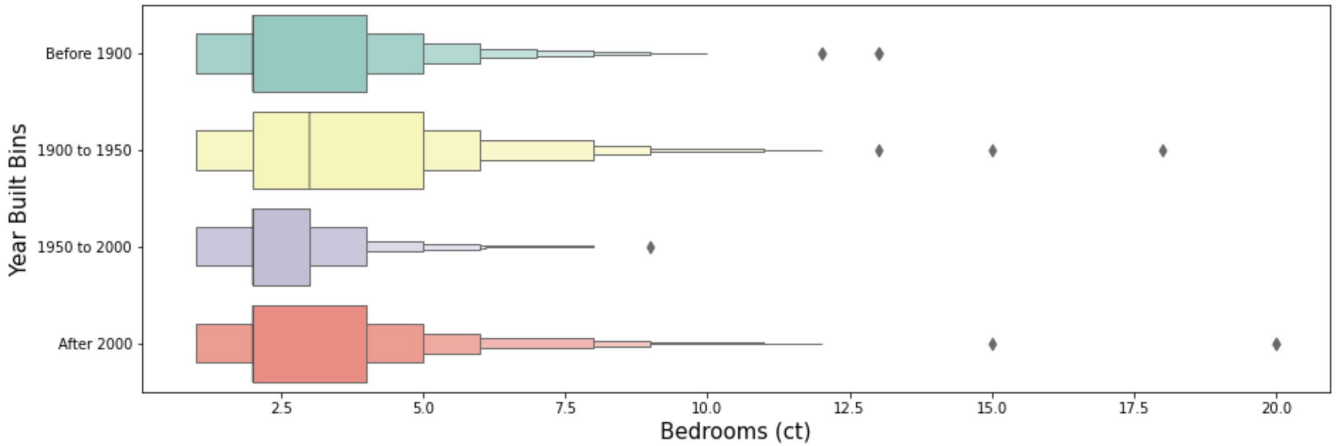


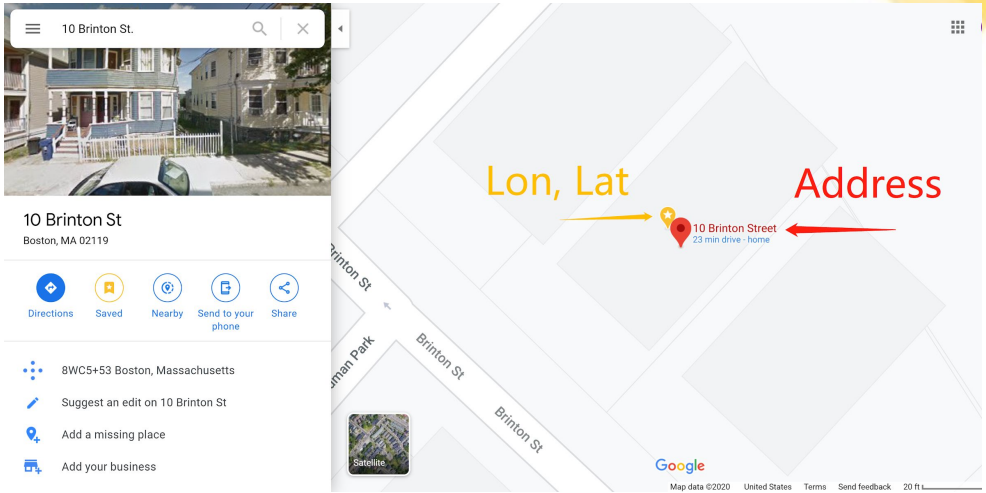
Fig.(above) House Built Year Histogram

Fig.(below) Bedroom number distribution in different Built Year



We have different levels of geo-data:

1. Exact-level:
   a. Address - 10 Brinton St.
   b. Longitude, Latitude
2. Zip Code-level: 02119
3. City-level: Boston
4. State-level: MA



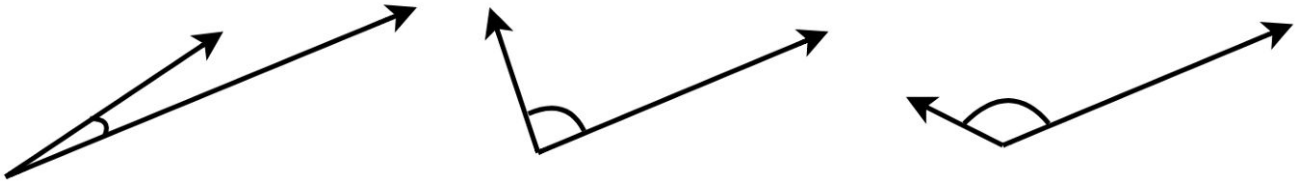
How we treat categorical features? E.g. cooling, property type.

Case study on **Cooling** columns:

1. Merge categories based on (subjective) domain knowledge
2. Apply one-hot encoding
    a. User case: prefer no central AC

Baseline Model

- Data Preprocessing
  - Normalization of Features
- Model
  - Content-Based Filtering
  - Similarity Metric - Cosine Similarity



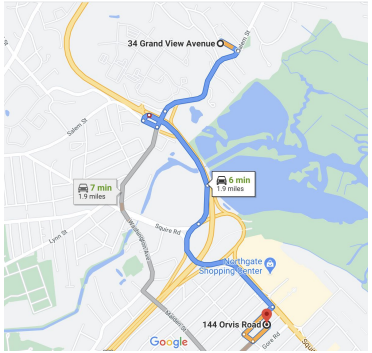
Small angle
Cosine Similarity ≈ 1

Near perpendicular
Cosine Similarity ≈ 0

Opposite directions
Cosine Similarity ≈ -0.8

Use Case Validation: What homes are similar to 56683068?

	cc_list_id	original_list_date	current_list_price	current_status
1	56683068	2020-09-04	474900.0	4
0	145525704	2020-07-07	474900.0	4
5405	56683877	2020-02-26	449900.0	6
5404	137981366	2020-04-17	449900.0	1
3619	56683070	2020-02-07	469900.0	6



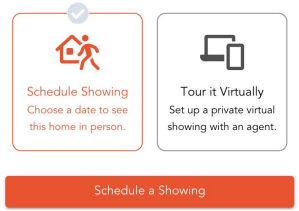
144 Orvis Rd (1954) - 1486 sqft
145525704
56683877

34 Grand View Ave (1960) - 1638 sqft
56683877
137981366

134 Orvis Rd (1950) - 1080 sqft
56683070

Future Work

- Improve the baseline model with information in text and images
- Explore user data from REX website



- Use user data to improve and validate our recommendation model
- Potential user testing using AWS Mechanical Turk (suggested by Andy)
- Integrate our model into REX’s current system

Project Timeline

Project Setting + M1

- Understand the features in selecting an ideal house
- **Define** the **top k primary features** that affect customer house selection

M2: advanced model

- Develop a **CF model** that can recommend housing using **features** that can be extracted from **REX's current models**.
- **Update** the model **using feedback** from the customer.

M3: build our model

- **Extract** our own **features**: image processing and NLP
- **Add** in new features to the **CF model** from M2 (hopes an improvement quality)
- [Stretch Goal] If time permit, plan for **model deployment with REX**

M3: wrap-up + Project Summary

- Offer **final deliveries**
- Prepare final report and presentations
- Collect **feedbacks** from REX
- Code review