# Digital Real Estate Index

Statement of Work (SoW) Document

**Prepared by:** Jessica Wijaya, Will Fried, Shucheng Yan, YiXuan Di
**Advisor:** Zona Kostic
**Prepared for:** Andy Terrel, aterrel@rexhomes.com

## Background

REX is a real estate technology company that is working to disrupt an industry that hasn't seen much innovation over the past 50+ years. REX uses machine learning and Advertising Technology to avoid using the multiple listing service (MLS) and directly target buyers interested in REX listings. By bypassing that MLS, which obligates the seller to pay a 5-6% commission fee to the seller and buyer agent, and by harnessing technology to streamline the real estate workflow, REX is able to reduce the cost of the home transaction to as low as 2%.

Beyond building out its core services, REX is interested in applying data science to real estate domains that don't currently utilize advanced statistical or computational tools. One such example is real estate indices (REIs), which seek to measure the overall health of real estate markets ranging from the national to the municipal level. Two of the most prominent REIs are the Case-Shiller Home Price Index and the National Association of Realtors (NAR) Confidence Index. These REIs are consulted by a wide range of stakeholders including homebuyers and homeowners, investors, developers and brokerage firms to better understand current market conditions and trends.

While existing REIs are trusted to paint a picture of the current real estate environment and provide context using historical data, they are not designed to predict how real estate markets will evolve in the future. This is a major deficiency as all the stakeholders listed above have a strong interest in understanding how market conditions will progress over time. Therefore, the goal of our project is to identify both digital and traditional indicators that can predict market conditions and integrate them into a predictive model, which REX can then display to its customers and the industry as a whole.

## Problem Statement

The goal of our project is to build a model that can predict *digital* real estate index as a measure of the overall health of the real estate market, which relies on both non-traditional (*digital*) and traditional data. This is different from the currently existing indices (such as NARS and Case-Shiller Home Price Index) in that the current indices only reflect the real estate market from the traditional view, while we are trying to incorporate non-traditional data such as real-time bidding for digital marketing ads, online listing, etc. In addition, while the existing indices reflect the current market condition from a high level standpoint (i.e. national level, market area level), we want to develop a more accurate and localized prediction (i.e. digital index for each neighborhood/zipcode) to give a more nuanced picture of where the real estate market is headed.

REX is interesting in this project because it believes that there are a lot of insights we can dig from the consumer behavior on the web that are not captured by existing real estate indices. For these reasons, we are going to do more research on this digital data, blending it with the traditional data and indices, to create a more accurate prediction for the real estate market.

While we don't yet have access to the digital datasets, there are several ways in which we think they may help us predict market conditions. Regarding the online listings dataset, the more listings that are on the market at a given time, the more sales are likely to occur over the following several months. The real time bidding dataset may also provide many insights. Because buyers and sellers often spend months researching and making preparations to move before they contact a buyer agent or list their home, their online activity may signal that they are preparing to enter the market. For instance, a prospective buyer may spend hours exploring different neighborhoods on Zillow, while a prospective seller may visit home improvement websites to fix up their home before putting it on the market. This information is recorded in the cookies that follow them around the web and are shared with real time bidding platforms. Meanwhile, if advertisers are willing to bid higher on visitors to Zillow in particular regions, this may indicate that market conditions are expected to be particularly strong in certain areas.

**Resources Available**

There are 2 main non-traditional datasets (provided by REX) that we are going to use:

1. Data on Real Time Bidding for marketing display ads (from Beeswax). Information from these bids include name of the winning bidder/company, dates of the bid, price, etc.
2. Online listing crawls from MLS aggregations. Information for the online listing include the address of the building/house, transaction price, transaction dates, etc.

On the other side, we are also going to rely on these traditional data sources:

1. Demographic Data (consumer habits, voter files, census data)
2. First American Assessor Data
3. First American Deed and Mortgage Records

**High Level Project Dates**

Our plan is to first understand the data a bit more. Since these data has never been processed before by REX, we need to do more research and EDA to figure out which features are important, which areas we need to dig more, how to connect them all, and what data (if any) we need to obtain further. The data itself is massive (in TeraByte at minimum), and hence the exploration will take more time. Our game plan currently is outlined below:

- Exploration on the digital data
- Research on the real estate market at each market area, and decide which to focus on
- Draw connections between digital (non-traditional) and traditional data
- Make predictions for digital real-estate index for the particular market area
- Generalized predictions for other market areas (if possible)

**Project Timeline**

| Fall Ending | Tentative Milestone or Goal |
|---|---|
| Sept 18, 2020 | Project Setup:<br>- Set-up Git repository<br>- Team communication channel selected: Slack; TF added<br>- Set up meeting times with Chris, TF, and team |
| Sept 30, 2020 | Milestone 1<br>- First draft on EDA:<br>  - Decide which market area to focus on<br>  - Exploration on beeswax data, perform feature engineering/feature selection<br>- All files on Github<br>- Circle back with REX for an update<br>- Draft first version of report |
| Oct 21, 2020 | - Update EDA<br>- Draw connections between datasets (digital and traditional data)<br>- Prepare presentation for Milestone 2 |
| Oct 28, 2020 | Milestone 2:<br>- Update EDA (if any)<br>- Create a baseline model for predictions<br>- Circle back with REX for an update<br>- Update report |
| Nov 18, 2020 | - Refine model<br>- Evaluate model by backtesting on historical data |
| Nov 25, 2020 | Milestone 3:<br>- Finalize model<br>- Finalize report<br>- Start preparing for final presentation |
| Dec 16, 2020 | Final Presentation |