# Modern Web Scraping and Data Analysis Tools to Discover Historic Real Estate Development Opportunities

By

**Nile Berry**

Bachelor of Public Policy
Hamilton College, 2014

Submitted to the Program in Real Estate Development in Conjunction with the Center for Real Estate in Partial Fulfillment of the Requirements for the Degree of Master of Science in Real Estate Development

**at the**

**Massachusetts Institute of Technology**

**September 2022**



**Signature of Author**_____
**MIT Center For Real Estate**
**July 28, 2022**

**Certified by**_____
**James Scott**
**Research Scientist, Center for Real Estate**
**Thesis Supervisor**

**Accepted by**_____
**Professor Siqi Zheng**
**Samuel Tak Lee Professor of Urban and RE Sustainability**
**Department of Urban Studies and Planning**
**Faculty Director, MIT Center for Real Estate**

# Modern Web Scraping and Data Analysis Tools to Discover Historic Real Estate Development Opportunities

By Nile Berry

Submitted to the Program in Real Estate Development on July 28th, 2022 in Fulfillment of the Requirements for the Degree of Master of Science in Real Estate Development

**ABSTRACT**

The National Parks Service (NPS) is responsible for a database that catalogs every nationally recognized historic building and historic district across the United States. If listed as historic, the property can qualify for both state and federal historic tax credits which can subsidize up to 45% of the Qualified Rehabilitation Expenses (QREs) of the project, depending on the specific state. These tax incentives can significantly increase the financial return profile of these redevelopment projects. However, finding these qualifying sites across the United States is challenging given the size of the NPS database. With over 97,000 rows of static data, the NPS database is unwieldy and difficult to maneuver. Moreover, there is no way to proactively use the tool to find acquisition opportunities.

This thesis project aims to solve this problem by creating an acquisition analytics funnel that aggregates data from multiple online sources and layers it to create a dynamic way to source new historic redevelopment projects. The initial focus area of the thesis is the state of Maine and the subject of the thesis is *Historic Tax Credit View* (*HTC View),* a digital data analytics platform conceived built and owned by the author. The platform combines the NPS database with automated web-scraping algorithms to parse publicly available census and market demand data that indicate whether certain markets are of higher investment value than others. Through the development of *HTC View*, the author and outside partners have raised funds to make a purchase of a historically recognized former Milling Site in Skowhegan, Maine that was originally identified as an opportunity by the platform. The results of this research demonstrate the effectiveness of adopting web-scraping technologies and the usefulness overall to real estate development professionals.

*The HTC View platform can be accessed at https://htcview.vertifyanalytics.com/ using login details that may be provided upon request via nberry@mit.edu*

**Thesis Supervisor: James Scott**

**Title: Research Scientist, Center for Real Estate**

## ACKNOWLEDGEMENTS

Want to thank the entire MIT community and Center for Real Estate for being such a fantastic place to learn and develop an area of expertise. A particular thank you to James Scott for his consistent wise words and guidance on shaping this thesis research.

## IMAGES & TABLES

## LIST OF FIGURES + IMAGES

# CHAPTER 1: INTRODUCTION

## 1.1 HISTORY OF HISTORICAL TAX CREDITS

The federal historic tax credit program, instituted by Congress under the Tax Reform Act of 1986, provides a 20% credit for the rehabilitation of qualified historical buildings. In the 36 years of activity, the program has been heralded as a widespread success, touting extensive benefits to the national economy generating more than an estimated $195.2 billion in GDP since 1978. In 2020 specifically, the Federal Historic Preservation Tax Incentives Program contributed more than $13.8 billion in output in terms of goods and services to the U.S. economy, generated approximately 122,000 jobs, and added an overall $7 billion in gross domestic product (GDP)[1].

The recognized success of the federal program has spurred many US states to create a version of the historical tax credit that works in conjunction with the federal credit. Today, 39 of 50 states have some version of a state credit that works in parallel with the federal credit. In the state of Maine, for example, the state government offers developers an uncapped[2] 25% credit on qualified rehabilitation expenses which, when combined with the 20% federal credit, covers 45% of every dollar spent to rehabilitate a historic building. Through both the state and federal tax credit, these historic redevelopments can have an extremely attractive investment return profile, particularly when the project costs are weighted more towards the redevelopment of the site, versus the acquisition of the property.

The National Parks Service (NPS), the same federal organization that manages all the US National Parks, manages the public database that categorizes all the buildings and districts that qualify for these historic credits, which is updated on a bi-annual basis with newly qualifying buildings and districts. As of this writing, the database contains 97,225 rows of data across both historic buildings, districts, with locations in nearly every city in the United States downloadable in Microsoft Excel. The existence of the database in itself is a successful feat of government

[1] Rutgers University's Center for Urban Policy Research, https://www.nps.gov/orgs/1207/htc-economic-report-2020.htm

[2] Uncapped means that there is no limited on the dollar value across projects that can be given historical tax credits in a given fiscal year

accounting, however the size alone of the data makes it extremely unwieldy to navigate from a public perspective. For the most part, the database today is a reactive tool used by prospective investors and historic consultants to confirm whether a project is listed by NPS. The aim of this thesis and HTC View is to flip the script and make this database a proactive, investment discovery and acquisitions tool.

Historical properties are listed both individually by NPS and as part of a district. A historical district is generally defined by NPS as a geographic boundary in a city in which all buildings within that defined area are considered historic.

Once a developer acquires a property eligible for historic tax credits, all of their Qualified Rehabilitation Expenditures (QREs) are eligible for the 20% federal credit and up to an additional 25% state credit. A QRE is broadly defined by NPS as any “renovation, restoration, or reconstruction of a building, but does not include an enlargement or new construction.” These QREs cover nearly all traditional ‘hard costs’ associated with redevelopment along with a variety of soft costs like: architectural and engineering fees, and ‘reasonable’ developer fees.

Once a historical tax credit has been granted by NPS after a project’s completed rehabilitation, that credit amount can be claimed against an individual’s income tax for up to 20 years. For example, suppose the full rehabilitation costs of a historic project was $4,000,000 in the state of Maine where there is a 25% credit offered in addition to the 20% federal credit. In this scenario, one would receive a credit of $1,800,000. This $1,800,000 could then be claimed against personal income taxes in a single calendar year, or across multiple to meet the varying degree of one’s tax liability. Often, these tax credits are sold by the project developer at a discount upon the completion of the project to individuals or organizations that have disproportionately high-income tax burdens.

## 1.2 PROBLEM STATEMENT

Data analysis technology in the real estate industry has significantly trailed its peers, particularly when compared to public and private finance, technology, and

other popular investment asset classes (Mohanram 2020)[3]. The core reason for this appears to be that the underlying data for real estate is generally fragmented, siloed, inconsistent, and difficult to access. At the base level, real estate data tends to rely on specific manual inputs, which inherently creates a more precarious foundational structure for measuring broader macro trends (Mohanram 2020). New aggregating tools on the internet, however, are changing this dynamic, by creating rigid systems for the ways in which big data is uploaded to the web and accessed by the public. These new avenues of data collection create the opportunity to build automated systems that can, in real time, provide accurate snapshots of aggregated market dynamics that previously required extensive manual work. This thesis aims to use these publicly available real estate data tools to identify historically eligible projects in markets that indicate a potential for a higher degree of investment success. Furthermore, this thesis illustrates that the current methods used in real estate to source and discover potential projects are highly antiquated and poised to be disrupted through the use of advanced data aggregation technologies that can reliably collect, analyze, and serve users with unobjective insights into current investment market conditions.

As a real-world example, this thesis focuses on *HTC View,* a proprietary data analytics platform designed and developed by the author. The platform demonstrates how web-scraping and, more specifically, HTML-parsing algorithms[4] can be used to collect data from a variety of sources that can be combined to assist real estate development professionals in the search for new sites. *HTC View* was designed on the premise that it could complete work that would otherwise take a team of analysts hundreds (if not thousands) of hours to collect manually. Moreover, the data that is collected is aggregated and displayed for users convenience and also stored in order to develop key trend analysis over time. These features simply are not available on most public real estate aggregators like Zillow and Apartments.com. *HTC View*, and similar web scraping tools can be an invaluable investment discovery tool for real estate professionals,

[3] MOHANRAM, P. S. 2020. A Brave New World: The Use of Non-traditional Information in Capital Markets. *World Scientific Book Chapters*, 217-237

[4] Web scraping defines a broad range of operations of which HTML parsing is one. HTML parsing is an operation that deals specifically with extracting data listed on public HTML coded webpages

particularly those working in smaller teams that need benefits more significantly from the automation of mundane, repeatable, tasks.

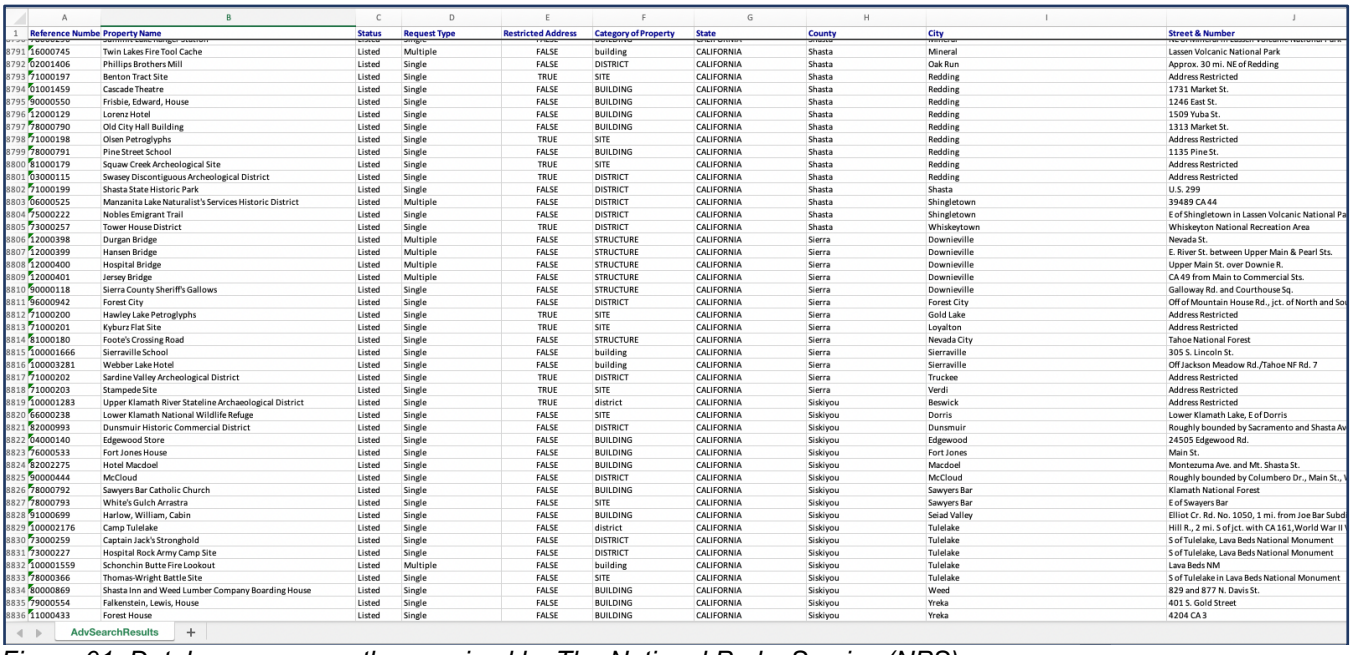
The focus area of this research is the United States and specifically the National Historic Places Registry database (maintained by the NPS) when combined with live sub-market data in every major US city. The condition and methodology of the NPS Historic Registry is emblematic of the broader problem within real estate data collection and accessibility. Specifically, the NPS registry relies on assembly of handwritten documents created by property owners and historical consultants created as part of the application to qualify that specific site for federal historic tax credits. These physical documents are manually scanned into the National Parks Service database and then manually cataloged into a .csv data table. Today, this process has been completed (at least) 97,000+ times by representatives of the National Parks Service. As new properties and sites are added each year, this database will become increasingly difficult to navigate and work with, particularly for functions of new property discovery. The proposed solution to this is the *HTC View* platform. *HTC View (i.e Historical Tax Credit View)* is a web-scraping platform that combines this static NPS database with live sub-market data (from Zillow's API) to create a comprehensive, sortable, view of the viable historic tax credit eligible properties, combined with that sub-market's live market characteristics. As new historical properties are added each year, the platform is designed to keep the National Registry document up to date, while the automatic links to public market data is designed to update daily to allow for a user to track and monitor trends over time. Each day *HTC View* exists, the more useful it becomes to a user to monitor and track market dynamics and trends.

To the author's knowledge, there are no available data aggregation tools that leverage the data used by National Registry and *HTC View* represents a new software platform solving a unique challenge for developers of historically eligible properties. Data feeds that return detailed sub-market information, on the other hand, are widely accessible and can come from a variety of website sources[5]. The core problem across all these sub-market data sources is that they are generally working with different sets of listing data  at the foundational level given

[5] Examples of live sub-market website data sources include: Zillow, Apartments.com, Costar, Loopnet, Crexi, CRS Data, among others

access to unique segments of buyers. Based on manual inputs, these different manual inputs can return vastly different insights. In public finance, corporate data is methodically logged, regulated, and published in a mandated cadence for public consumption. In real estate, data is aggregated from the masses, consolidated on unsupervised platforms, and often incomplete. Real estate professionals have all become comfortable with this "best-estimate" approach, and work with the best that's available to judge and base their analysis (Winson-Geideman 2017)[6]. These public datasets there are often missing values, redundant, and are inconsistent depending on the source and what one is looking for. All of the submarket data for *HTC View* is collected by autonomous scripts, eliminating human oversight and reducing human error from the process. While the magnitude of market trends tends to differ across platforms, the direction of the trend tends to be more or less consistent.

For the purposes of this analysis, Zillow's market data API was selected as our single source to allow for the most consistent and reliable comparison as possible. The goal is to create an investment analysis tool that provides investors with a centralized and efficient way to evaluate real estate investments across hundreds of markets.



| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Reference Numbe | Property Name | Status | Request Type | Restricted Address | Category of Property | State | County | City | Street & Number |
| 8791 | 16000745 | Twin Lakes Fire Tool Cache | Listed | Multiple | FALSE | building | CALIFORNIA | Shasta | Mineral | Lassen Volcanic National Park |
| 8792 | 62001406 | Phillips Brothers Mill | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Shasta | Oak Run | Approx. 30 mi. NE of Redding |
| 8793 | 71000197 | Benton Tract Site | Listed | Single | TRUE | SITE | CALIFORNIA | Shasta | Redding | Address Restricted |
| 8794 | 01001459 | Cascade Theatre | Listed | Single | FALSE | BUILDING | CALIFORNIA | Shasta | Redding | 1731 Market St. |
| 8795 | 90000550 | Frisbie, Edward, House | Listed | Single | FALSE | BUILDING | CALIFORNIA | Shasta | Redding | 1246 East St. |
| 8796 | 72000129 | Lorenz Hotel | Listed | Single | FALSE | BUILDING | CALIFORNIA | Shasta | Redding | 1509 Yuba St. |
| 8797 | 78000790 | Old City Hall Building | Listed | Single | FALSE | BUILDING | CALIFORNIA | Shasta | Redding | 1313 Market St. |
| 8798 | 71000198 | Olsen Petroglyphs | Listed | Single | TRUE | SITE | CALIFORNIA | Shasta | Redding | Address Restricted |
| 8799 | 78000791 | Pine Street School | Listed | Single | FALSE | BUILDING | CALIFORNIA | Shasta | Redding | 1135 Pine St. |
| 8800 | 61000179 | Squaw Creek Archeological Site | Listed | Single | TRUE | SITE | CALIFORNIA | Shasta | Redding | Address Restricted |
| 8801 | 03000115 | Swasey Discontiguous Archeological District | Listed | Single | TRUE | DISTRICT | CALIFORNIA | Shasta | Redding | Address Restricted |
| 8802 | 71000199 | Shasta State Historic Park | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Shasta | Shasta | U.S. 299 |
| 8803 | 06000525 | Manzanita Lake Naturalist's Services Historic District | Listed | Multiple | FALSE | DISTRICT | CALIFORNIA | Shasta | Shingletown | 39489 CA 44 |
| 8804 | 75000222 | Nobles Emigrant Trail | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Shasta | Shingletown | E of Shingletown in Lassen Volcanic National Pa |
| 8805 | 73000257 | Tower House District | Listed | Single | TRUE | DISTRICT | CALIFORNIA | Shasta | Whiskeytown | Whiskeytown National Recreation Area |
| 8806 | 12000398 | Durgan Bridge | Listed | Multiple | FALSE | STRUCTURE | CALIFORNIA | Sierra | Downieville | Nevada St. |
| 8807 | 12000399 | Hansen Bridge | Listed | Multiple | FALSE | STRUCTURE | CALIFORNIA | Sierra | Downieville | E. River St. between Upper Main & Pearl Sts. |
| 8808 | 12000400 | Hospital Bridge | Listed | Multiple | FALSE | STRUCTURE | CALIFORNIA | Sierra | Downieville | Upper Main St. over Downie R. |
| 8809 | 12000401 | Jersey Bridge | Listed | Multiple | FALSE | STRUCTURE | CALIFORNIA | Sierra | Downieville | CA 49 from Main to Commercial Sts. |
| 8810 | 90000118 | Sierra County Sheriff's Gallows | Listed | Single | FALSE | STRUCTURE | CALIFORNIA | Sierra | Downieville | Galloway Rd. and Courthouse Sq. |
| 8811 | 96000942 | Forest City | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Sierra | Forest City | Off of Mountain House Rd., jct. of North and So |
| 8812 | 71000200 | Hawley Lake Petroglyphs | Listed | Single | TRUE | SITE | CALIFORNIA | Sierra | Gold Lake | Address Restricted |
| 8813 | 71000201 | Kyburz Flat Site | Listed | Single | TRUE | SITE | CALIFORNIA | Sierra | Loyalton | Address Restricted |
| 8814 | 81000180 | Foote's Crossing Road | Listed | Single | FALSE | STRUCTURE | CALIFORNIA | Sierra | Nevada City | Tahoe National Forest |
| 8815 | 10001666 | Sierraville School | Listed | Single | FALSE | building | CALIFORNIA | Sierra | Sierraville | 305 S. Lincoln St. |
| 8816 | 10003281 | Webber Lake Hotel | Listed | Single | FALSE | building | CALIFORNIA | Sierra | Sierraville | Off Jackson Meadow Rd./Tahoe NF Rd. 7 |
| 8817 | 71000202 | Sardine Valley Archeological District | Listed | Single | TRUE | DISTRICT | CALIFORNIA | Sierra | Truckee | Address Restricted |
| 8818 | 71000203 | Stampede Site | Listed | Single | TRUE | SITE | CALIFORNIA | Sierra | Verdi | Address Restricted |
| 8819 | 100001283 | Upper Klamath River Stateline Archaeological District | Listed | Single | TRUE | district | CALIFORNIA | Siskiyou | Beswick | Address Restricted |
| 8820 | 66000238 | Lower Klamath National Wildlife Refuge | Listed | Single | FALSE | SITE | CALIFORNIA | Siskiyou | Dorris | Lower Klamath Lake, E of Dorris |
| 8821 | 82000993 | Dunsmuir Historic Commercial District | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Siskiyou | Dunsmuir | Roughly bounded by Sacramento and Shasta Av |
| 8822 | 94000140 | Edgewood Store | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Edgewood | 24505 Edgewood Rd. |
| 8823 | 76000533 | Fort Jones House | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Fort Jones | Main St. |
| 8824 | 82002275 | Hotel Macdoel | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Macdoel | Montezuma Ave. and Mt. Shasta St. |
| 8825 | 90000444 | McCloud | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Siskiyou | McCloud | Roughly bounded by Columbero Dr., Main St., |
| 8826 | 78000792 | Sawyers Bar Catholic Church | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Sawyers Bar | Klamath National Forest |
| 8827 | 78000793 | White's Gulch Arrastra | Listed | Single | FALSE | SITE | CALIFORNIA | Siskiyou | Sawyers Bar | E of Swayers Bar |
| 8828 | 91000699 | Harlow, William, Cabin | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Seiad Valley | Elliot Cr. Rd. No. 1050, 1 mi. from Joe Bar Subd |
| 8829 | 100002176 | Camp Tulelake | Listed | Single | FALSE | district | CALIFORNIA | Siskiyou | Tulelake | Hill R., 2 mi. S of jct. with CA 161, World War II |
| 8830 | 73000259 | Captain Jack's Stronghold | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Siskiyou | Tulelake | S of Tulelake, Lava Beds National Monument |
| 8831 | 73000227 | Hospital Rock Army Camp Site | Listed | Single | FALSE | DISTRICT | CALIFORNIA | Siskiyou | Tulelake | S of Tulelake, Lava Beds National Monument |
| 8832 | 100001559 | Schonchin Butte Fire Lookout | Listed | Multiple | FALSE | building | CALIFORNIA | Siskiyou | Tulelake | Lava Beds NM |
| 8833 | 78000366 | Thomas-Wright Battle Site | Listed | Single | FALSE | SITE | CALIFORNIA | Siskiyou | Tulelake | S of Tulelake in Lava Beds National Monument |
| 8834 | 66000869 | Shasta Inn and Weed Lumber Company Boarding House | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Weed | 829 and 877 N. Davis St. |
| 8835 | 79000554 | Falkenstein, Lewis, House | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Yreka | 401 S. Gold Street |
| 8836 | 71000433 | Forest House | Listed | Single | FALSE | BUILDING | CALIFORNIA | Siskiyou | Yreka | 4204 CA 3 |

AdvSearchResults

*Figure 01: Database as currently organized by The National Parks Service (NPS)*

[6] WINSON-GEIDEMAN, K., KRAUSE, A., LIPSCOMB, C. A. & EVANGELOPOULOS, N. 2017. *Real estate analysis in the information age: techniques for big data and statistical modeling, Routledge.*

## 1.3 RESEARCH AIM & OBJECTIVES

The specific objectives of the research are as follows:

1. To evaluate the technology and methodology for developing reliable web-scraping tools that return valuable investment insights for real estate professionals
2. To develop the specific data sets needed to most effectively benefit the work of real estate investment professionals looking for historical tax credit opportunities
3. To examine the quality of the output of *HTC View* using live data from the top 100 cities (by population) in state of Maine
4. To create a roadmap for future development and ways to improve the quality and functionality of the *HTC View* platform for real estate investment professionals

Web scraping is the process of setting up automated computer scripts with the purpose of retrieving data from the internet and pasting it into a database. Once added to a database this information can become much more valuable as it can be combined with many different sources and a variety of analytical functions can be performed. An Excel spreadsheet mimics the functionality of a database, however the underlying data is generally not connected to real-time figures and metrics.

For this project, we developed a variety of automatic scripts for gathering, updating, and cataloging real estate data across the selected data sources. Our analysis focused on pricing data most notably, followed by locational data of each property, demographic data, and time-stamps of when data assets were added and removed from our data feeds.

Another key goal for this project was to set up a framework that would catalog all the live data over any given period of time. Most popular real estate listing websites focus on giving the user a real-time snapshot of what's available in the market, but offer no functionality to look backwards to determine possible trends in financial market conditions. The *HTC View* web scraper is designed to run each day and append the information to a database that allows a user to observe key trends and shifts over any given period of time. Having a database of malleable

submarket real estate data has other research benefits like the ability to create new data fields that can observe other key insights into a market. For example, *HTC View* has a data field designed to track the amount of time that a rental property exists on Zillow public marketplace. Logging this value creates a key metric to determine a live market demand for different types of rental assets in that market which can then be extrapolated to similar submarkets.

The final piece of this research is to create a single, user friendly, web platform dashboard in which this data can be accessed by real estate investment professionals. This final web application *HTCView.com* will provide a team of analysts with all the necessary data analysis tools and interactive dashboards to focus their acquisition efforts and focus exclusively on the key submarkets for historic tax credit eligible opportunities.

### 1.4 SCOPE OF STUDY

Overall, this thesis aims to determine the specific functions that web-scraping technology can provide to investment professionals within real estate. Beyond this, this study aims to identify the key features of successful data scraping tools and ways platforms like *HTC View* could be built for other key objectives within the industry. The *HTC View* platform today is limited to the United States and relies on market data from Zillow's public API to gather key market insights on cities that have a disproportionate amount of properties that are eligible for federal historic tax credits, and also have a favorable state-run historical tax credit program.

### 1.5 HYPOTHESIS

As examined in the literature review, web-scraping tools can evaluate extensive amounts of disparate market data in a highly efficient way that can dramatically improve individual productivity. Despite this, these tools are seldom used in the real estate industry given the irregularity of the underlying datasets which are often reliant on manual inputs. Finally, a tool like *HTC View* can be an extremely valuable way to cut through the noise and funnel the locations that are most likely to provide viable sites for historic rehabilitation.

These observations lead to the hypothesis that:

*'Automated web-scraping and scripting algorithms can be a valuable tool for real estate development professionals, specifically with asset types that do not have existing reliable market data sources'*

## 1.6 THESIS STRUCTURE

The thesis is composed of the following Chapters:

**Chapter 1:** provides an introduction of historical tax credits and the problem

**Chapter 2:** a literature review exploring relevant work relating to the development of web scraping applications

**Chapter 3:** explores the product methodology and use case for HTC View

**Chapter 4:** highlights the existing HTC View functionality and codebase logic

**Chapter 5:** presents a key project identified with HTC View that has since been purchased and under development

**Chapter 6:** presents the conclusion of the product and the roadmap for future development

# CHAPTER 2: LITERATURE REVIEW

## 2.1 INTRODUCTION

A summary of the available literature explores and evaluates works that relate to big data and web-scraping within the real estate industry. The literature review also focuses on tools commonly used by developers to build web scraping tools and the ways in which they can provide unique value. Finally, the literature review analyzes some of the real estate industry backlash relating to web scraping technologies.

## 2.2 BIG DATA, WEB SCRAPING, AND THE REAL ESTATE INDUSTRY

Data is the oil of the digital era: an immensely, untapped asset that can be leveraged for outsized financial returns. (The Economist, 2017). At the heart of this opportunity is the innovation in which data can now be collected, stored, and then ultimately accessed.

The genesis of ‘Big Data’ is complex and attributed to a myriad of simultaneous technical improvements in computing, data analysis and data storage. The cheap costs of off-site cloud storage in particular have enabled a rapid adoption of big data strategies. Doug Laney, an analyst at the META Group (now Gartner) is one of the first researchers who identified the massive opportunity associated with the increasing volumes of data being generated globally. In a 2001 research report Laney theorized that there were three things that characterized the growth of this new class of information, making it markedly different from conventional data—“’Big Data’ is high-volume, high-velocity and high–variety information assets that demand cost-effective, innovative forms of processing for enhanced insight and decision making” (Laney, 2001).

As data scraping has grown in prominence the shortcomings of the method of analysis have also been brought to light. As Kimberly Winson-Geidman and Andy Krause argue that Big Data does not always provide better, more insightful results just because of its size. Rather, Big Data is prone to many of the same methodological problems smaller sets are, such as user input errors and sampling error. Big Data is also subject to issues not generally associated with

conventional data including source identification, particularly given the automated nature of some Big Data sources (Geidman-Krause 2017). Additionally, as it currently exists most Big Data is capturing mundane and obvious observations like pedestrian counts, mouse clicks, and credit card transactions (Kitchen 2014). These are often end results, not motivation and causes. The 'why' of any phenomenon is at the heart of any research question the relies on these big data techniques, which generally requires many overlapping datasets to determine colorations and causalities.

Another major shortcoming to big data is that it has also become clear that, taken out of context, Big Data can lose its meaning. Geideman and Krause further argue that real estate data is the most vulnerable to this issue as markets are very localized and if the information extrapolated from large data sets is often not reflective of local market conditions and thus meaningless (Geideman Krause 2017). Further, technical skills and understanding create barriers to accessing Big Data as well, as the process of extracting it requires knowledge of computer coding, database design, and some statistical analysis.

Within real estate, "big data" can be categorized in three distinct groups:

| Core | Static Spatial | Peripheral |
|---|---|---|
| Sale Transactions | Census information | Internet Searches |
| Lease Transactions | Road Network Data | Transit Boarding Data |
| Mortgage Information | Geographic information | Live Traffic Information |
| Tax Assessment Values | Aggregated Spatial Core Data | Point of Sale (POS) Data |
| Property Physical Characteristics | Urban Planning Forecasts | Geo-Located Tweets |
| REIT & Real Estate Stock Data | Spatial Economic Indicators | Pedestrian Foot Counts |

Organizing the data into these three distinct categories is helpful to conceptualize the types of data available and the various research questions that could be answered when various data sets are combined. For the most part, web scrapers within real estate are pulling in one or more of these types of data sets into a structured database, which allows a researcher to run specific analyses that allow them to gather a key analytical insight.

A key example of this is the use of Big Data in real estate to forecast housing prices. For example, Lynn Wu and Erik Brynjolfsson conducted a study in 2015 in which they analyzed Google search trends in the United States to foreshadow

housing price growth in specific markets. Their model outperformed the National Association of Realtors home price predictions by greater than 20%, highlighting the value of collecting real time “Peripheral” data to predict the “Core” data of real estate asset prices. Similar analyses have been conducted in the context of “Smart Cities” in which authors like Rob Kitchen demonstrated how the proliferation of smart phones, ubiquitous sensors, and improved wireless networks enable a new generation in urban planning and development. Further, this can data enable more efficient municipal resource allocation and city management (Kitchen 2013).

Overall, there are 2.5 quintillion bytes of data created globally each day. At this rate, the world is able to produce “90% of all the data ever created in just a two year time frame.” This scale of the data production almost ensures that it will be a significant part of our future. In real estate, this represents an enormous opportunity for the industry to use new variables to evaluate asset markets in completely new ways. The introduction of peripheral data represents a paradigm shift in knowledge and the types of questions researchers can begin to ask. Big Data, and its many applications, are undeniably important, but as with any source of analytical information the sources of this data must be rigorously examined and tested. After all, the output of any of these data models will only be as accurate as their inputs.

## 2.3 UNDERSTANDING WEB-SCRAPING MECHANICS AND PROGRAMS

The search and extraction of information from the World Wide Web (WWW) is usually performed by web crawlers which is simply a computer script that browses the Web in a preprogrammed manner. Web scrapers, however, are a type of Web crawler that not only can identify desired data, but then pull that data into a new desired format.

Online web-scraping tools are algorithms that are designed to parse information from data sources. They rely on simple ‘search and save’ algorithms that are at the core foundation of major search engine companies like Google. These are ‘string-searching’ algorithms that troll alphanumeric characters for pre-programmed matching keywords. To increase the accuracy of these scraping tools, it is recommended that programmers also utilize a process called DOM Parsing which

also specifies the location of the desired information on the web page. The specifics of the web page, dictates the type of web scraping that should be used in order to produce a successful return value for the program.

Web scrapers automate the search process on websites and the copy/paste function that a human would typically use to capture the same data. In summary, web scraping focuses on the identification of unstructured data on the Web, typically in HTML format, and transforms it into a structured database that can be analyzed and worked with in a more organized format.

There are a variety of methods to scrape online information ranging in complexity. The most basic is copy-paste which is a manual keyboard shortcut to quickly extract string data from a HTML webpage. The manual drawbacks of copy-paste have given way to more advanced techniques like: HTTP programming, DOM parsing, and HTML parsers which scrape HTML pages in unique ways. The usage of these methods (or combination of methods) is dependent on specific factors to that webpage. These may include: whether the page requires user authentication to access, the page location of the data shifting on the page, and irregularities in the format that the desired data is published.

There are a variety of publicly available tools, most of which have existed for over a decade, that help users build effective web scrapers to suit their specific data acquisition needs. While these tools have been around for some time, they are continuously growing in functionality and popularity as web scraping applications continue to grow in popularity.

**Beautiful soup (initial release 2004)**

Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

**Selenium (initial release 2004)**

Selenium is an open source umbrella project for a range of tools and libraries aimed at supporting browser automation. It provides a playback tool for authoring functional tests across most modern web browsers, without the need to learn a test scripting language.

**Import.io (initial release 2012)**

Import.io is a SaaS web data platform. It is an example of one of the for-profit companies that have developed a suite of data scraping tools that apply to many industry applications. It is much more user friendly than Beautiful Soup and Selenium but provides less flexibility in terms of programmability.

As stated previously, data scraping applications in the real estate industry have significantly lagged other major assets classes. This is partially due to the unreliability and irregularity of the underlying real estate data, but also in part to the limited avenues that exist to access the data (Snell 2013)[7]. The Costar Group, for example, is a $25bn company that operates one of the largest real estate data products within commercial real estate. Their business model involves the operation of a variety of different real estate listing marketplaces and then use that data to inform their broader data product. In a sense, the company effectively scrapes the data from their own marketplace products to inform macro insights. Costar has been accused of being a monopoly, an often-litigious organization to any competitive data product that mimics their data offering.

## CHAPTER 3: PRODUCT METHODOLOGY AND USE CASE FOR HTC VIEW

### 3.1 INTRODUCTION

This chapter describes the scraping technology used to retrieve, process and store the data from the NPS database and other online sources. The descriptions provide insight into how the *HTC View* platform works and are intended to provide a recommended framework for building similar applications.

### 3.2 DESIRED DATA

The core data that makes this analysis possible is the existence of the National Parks Service Register of Historic Places. This is a static document available to the public that is updated every 4-6 months, depending on the number of newly approved historic sites that they receive. The site data in this list is then run through a geotagging algorithm to give each site a latitude and longitudinal

[7] SNELL, J 2013. Use of Online Data in the Big Data Era: Legal Issues Raised By the Use of Web Crawling and Scraping Tools for Analytics Purposes

coordinates. Finally, this historic data is paired with real time market level data from Zillow that shows the real time rental attributes of that target market. Manually sifting through these various historical properties, alongside real time market level data would be an extremely time intensive task if taken on without the help of modern web scraping tools.

## IDENTIFYING HTC ELIGIBLE PROPERTIES USING WEB SCRAPED DATA

There are a variety of ways that real estate professionals identify and secure investment opportunities. Most of these methods are highly manual, relationship based, and require some degree of knowledge about the local market. For the purpose of this study, our focus is on properties that qualify for federal historic tax credits that also fall within a market that possesses key positive trends that would suggest a successful investment.

## DEFINING 'TARGETS OF INTEREST'

By having access to the NPS Historical Registry, we are in possession of all possible targets for our historical tax credit investment strategy. Thus, *HTC View's* purview of possible investment targets is not limited to the buildings simply for sale at any given time in the market. This provides real estate investors with a unique advantage in identifying and sourcing new investment opportunities in key markets, creating an opportunity to generate unsolicited sales offers to possible sellers.

For our analysis, investment "targets of interest" are specifically those that are eligible for federal historic tax credits in states that also have robust historical tax credit programs of their own that can be used in tandem. Furthermore, our analysis is specifically targeted at areas of consistent "Rent Per Square Foot" growth and fall within a specific geographic radius of the central downtown business district.

## 3.3 DATA SOURCES

Each desired city has a unique Zillow website URL which are the data sources and 'scraping targets'. The uniformity of the Zillow website structure allows for a single algorithm to perform the same scraping sequence for each city, every day to generate uniform data for the platform.

### 3.4 DATA STORAGE

All of the retrieved data is stored on Amazon's Simple Storage Service (AS3) cloud servers. The private server space is referred to as a 'bucket' and it can be easily integrated into the platform via Amazon's Application Programming Interface (API). Amazon's AS3 is the most widely used cloud storage service in the world.

### 3.5 HTC VIEW PROGRAM ARCHITECTURE

Figure 3.1 diagrammatically illustrates the general architecture in which *HTC View* was constructed, is accessed by the program, and then ultimately stored within a SQL database.

As noted above, each LGA website has a unique URL and structure which requires a different algorithm for each website. A typical REGEX pattern that will locate any user-generated keyword in a PDF document will be in the form:

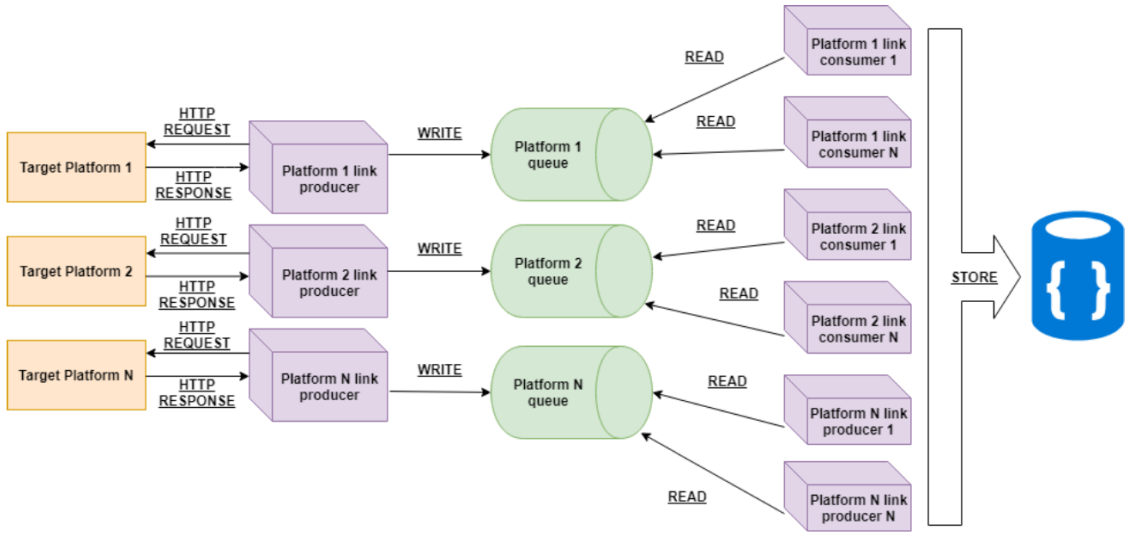



*Figure 3.1: Architecture view of HTC View data connectivity*

## CHAPTER 4: THE HTC VIEW ONLINE PLATFORM

### 4.1 INTRODUCTION

This chapter provides an overview of the key technologies and modules that were used to build the *HTC View* platform. Furthermore, the layout, structure and User Interface (UI) of the platform is discussed and provided as reference.

### 4.2 TECHNOLOGY STACK

The technology stack used to build the *HTC View* leverages various code modules and plugins. These technologies are summarized below and grouped into front-end (user facing), back-end, and the database structure. In addition, there are a number of sample images of the user interface with labeled tags highlighting the function of each field.

Existing Tech Stack:

**FRONT END:** Livewire and CSS Bootstrap for design of the data table

**BACKEND:** PHP/Laravel 7

**DATABASE:** MySQL

**PLUGINS:** RapidAPI: Zillow Live Data

## 4.3 KEY FUNCTIONALITY OVERVIEW

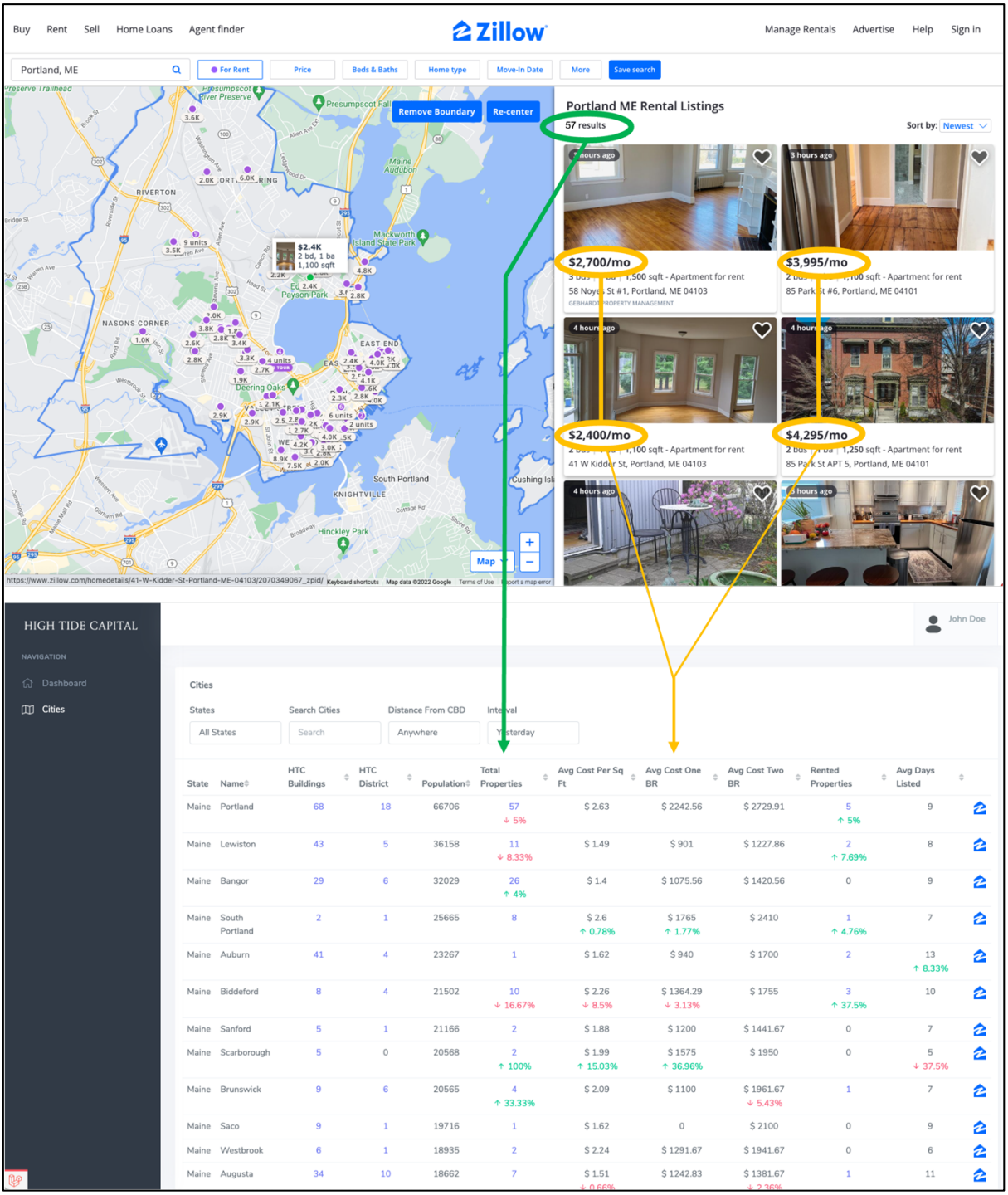



*Figure 4.1 Basic Illustration of how web scrapers work on Zillow.com to extract data to the HTC a database*

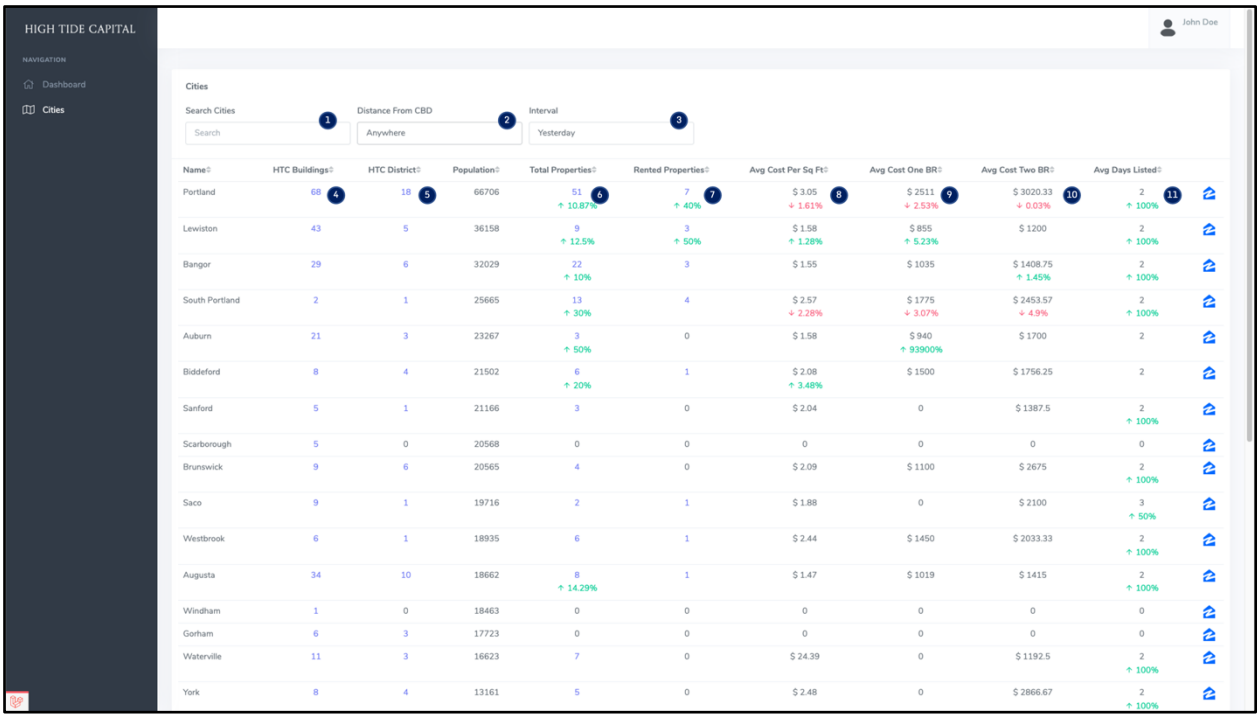



*Figure 4.2: Main data dashboard of HTC View, filtered by top 100 populated cities in Maine*

- **#1: Search Cities:** Main search bar for users to toggle between cities. For testing purposes, the platform today only has the top 100 most populated cities in the state of Maine, but that will be slowly expanded to all 39 states with a state-run historical tax credit program.
- **#2: Distance From the CBD:** A sort function for all data on the platform by the distance from the central business district. Distance from the city center has proved to be a reliable sorting function to find buildings of the desired investment size that have the highest ROI to convert into multifamily apartments.
- **#3: Time Interval:** A sort function allowing the user to modify the displayed data over any period of time that they select. The percentages below each figure indicates the difference between the current value and the inputted time interval by the user. Each day, as the platform continues to add more data, these sort functions will become more robust.
- **#4: HTC Buildings:** A static data feed pulling all the historic buildings listed by the NPS in that city. This static data is updated in six month intervals when the NPS has indicated that they publish new historical sites
- **#5: HTC Districts:** A static data feed pulling all the historic districts listed by the NPS in that city. Akin to HTC Buildings, this data feed is updated in

six month intervals when NPS indicates a new publishing of the National Register of Historic Places

- **#6: Total Properties:** This is a snapshot in time showing all the available properties on Zillow in that specific market. This number fluctuates daily as listings are added and removed from the inventory of each city. Are these properties for sale or for rent?
- **#7: (Recently?) Rented Properties:** This is a data set created by comparing the Total Properties scrap from one day to the next. Marking each property with a unique Zillow reference ID allows the program to see which apartments have been taken out of circulation from one day to the next. Once a listing is removed from inventory, the program calculates the difference in days between when it was posted and when it was removed to create a “Days Listed Value”. This “Days Listed” value is then averaged across all properties removed from circulation over the specified time. This is a valuable dataset as it indicates the rental demand velocity of a certain market. Observing this demand trend overtime is a key insight to a submarkets demand for new inventory.
- **#8: Average Cost Per Square Foot:** This is a simple calculation looking across all available and recently rented inventory over a specified period of time. The calculation simply divides total living space by gross monthly rent to create a apples to apples value that allows a user to easily compare costs across markets.
- **#9: Average Cost Per One Bedroom:** Simple average of all the available (and rented) one-bedroom apartments in that submarket
- **#10: Average Cost Per Two Bedroom:** Simple average of all the available (and rented) two-bedroom apartments in that submarket
- **#11: Average Days Listed:** Takes the average of all rented properties “Days Listed” to display a single numerical value. This average value is a key method for measuring demand velocity in a market over a set period of time.

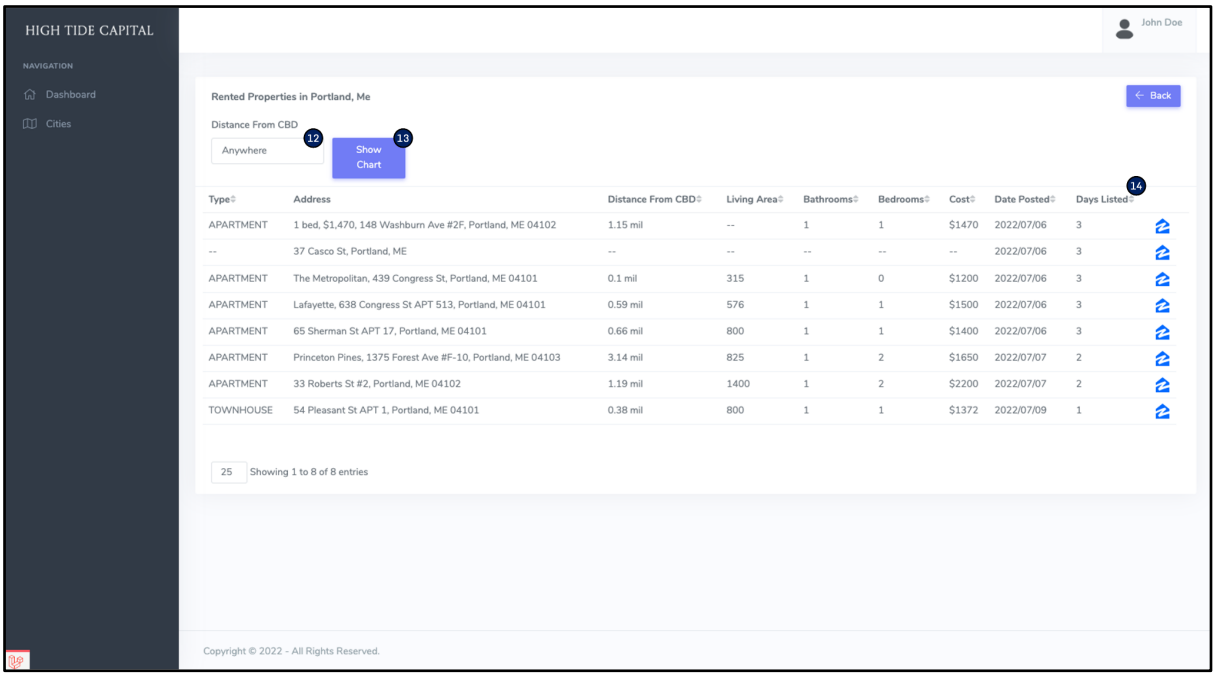



*Figure 4.3: Subpage view of the rented properties in a given "Cities"*

- **#12: Distance from CBD:** Main search bar for users to zero in on rented properties within a specific geographic radius of the central business district. This is a key feature as proximity to city center general coincides with other key investment factors that are typically harder to measure like building size and unit density.
- **#13: Days Listed:** A data field that is returning the difference between when a listing was published on the Zillow platform and when it was taken down. This serves as a high-quality metric to assess the market demand in a given market and the type of rentals (and neighborhoods) that are most popular.
- **#14: Show Chart:** A visual representation of the table view that shows property monthly rent on the y-axis along with the data of rental on the x-axis. Doing so creates a visual trendline for a given market indicating consumer leasing demand.

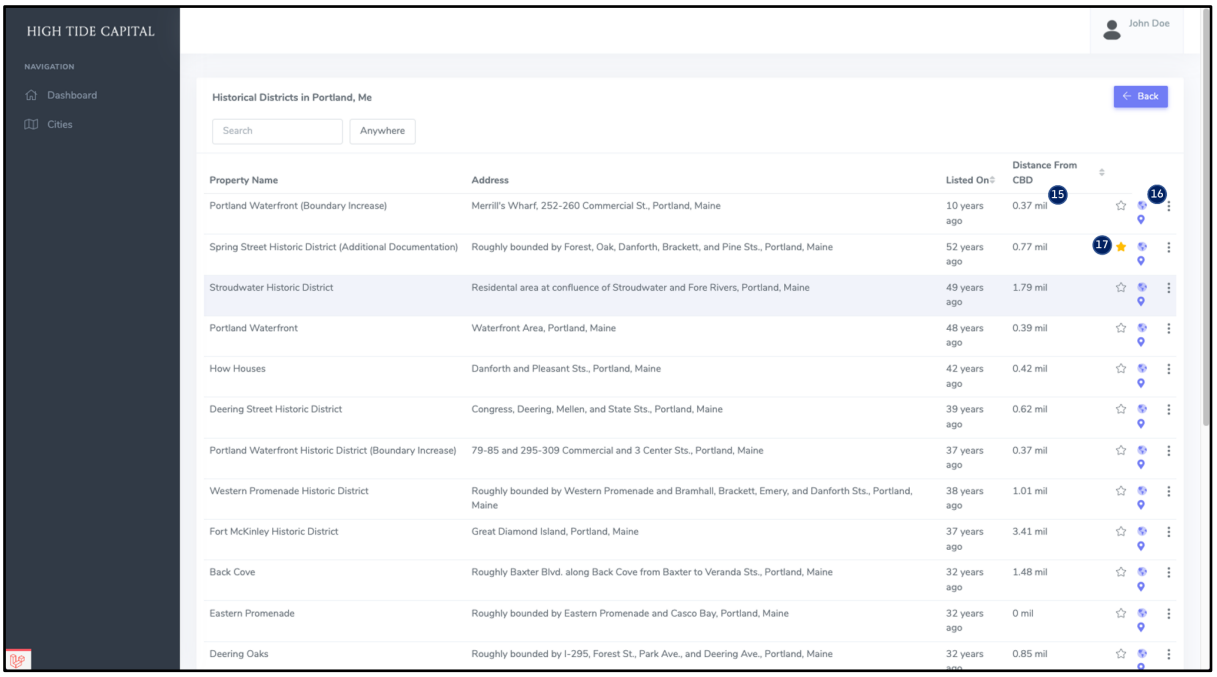



*Figure 4.4: Subpage view of the historic districts and their proximity to the CBD in Portland Maine*

- **#15: Distance from CBD:** Assigned lat/long points for each district allow the user to see districts closet to the central business district (CBD) of that city. As stated, closer proximity to the city center is highly correlated with other desired investment qualities.
- **#16: Google Earth Embed:** An embedded Google Earth link that allows a user to quickly navigate to that specific historic district or building to view in either Google Streetview or in satellite images.
- **#17: Bookmarking Tool:** *HTC View* was designed for teams to work together and conduct analysis on markets in real time. The bookmarking tool is a key feature that allows users to build on the work of their colleagues to see highlighted historic buildings and districts.

### 4.4 Future Work + Functionality

Overall, the development of *HTC View* proved successful in creating a data aggregation funnel to assist in the discovery and ultimate acquisition of properties eligible for historic tax credits. Today, this platform is used internally by an investment group as one of the many tools to guide their investment decision making across US markets.

The platform provides methods for various combined searches to quickly identify key markets and also key assets within those markets that are most suited as a historical redevelopment. As this platform continues to scrape data, the insights generated will become more useful, particularly as trends of market data are able to be displayed over longer periods of time and can endure the effects of market cycles.

There is significant future functionality scheduled for *HTC View*, most notably the ability to layer over commercial sales data, property records, and building total square footage estimates. Additional functionality on the platform also includes the ability for a user to build their own queries (with the data available) and have the results display directly on the shared dashboard. For example, a user will be able to use the scripting function to begin crawling the data on *HTC View* for specific attributes (i.e. Xs consecutive weeks of reduced “Average Days Listed” combined with Xs consecutive weeks of positive rent growth, etc).

## CHAPTER 5: PLATFORM OUTPUT + FINDINGS + THE SKOWHEGAN SPINNING MILL

### 5.1 INTRODUCTION

In this chapter, we evaluate the analysis of HTC View and present the use cases in which the program can be used by real estate investment professionals. In one such case, an investment group was able to use the platform to identify a high-value market which ultimately led to the purchase and renovation of the Spinning Mill, a 40,000 square foot residential mill conversion located in Skowhegan, Maine.

### 5.2 VALUE OF FINDINGS + RESULTS

As stated previously, the insights generated from the platform improve over time as more data is collected and analyzed by the HTC View program. At its core, the program is aggregating thousands of points of data to make it a dynamic, easily searchable tool for the user. Moreover, the program creates new fields of data that indicate specific sub-market multi-family rental performance. One key field of data generated from this program is the “Average Days Listed” column, which calculates the duration of time (in days) between when a listing is posted on

Zillow and then removed. This figure is then averaged across every property available/no longer available to return a single average that applies to the entire submarket. This single variable allows investors to determine both supply and demand in the same figure.

This variable was a key piece of data to find Skowhegan, Maine, and when combined with other factors (i.e. Average 1BR Rent, Distance from the CBD, etc) was clear that the market satisfied a number of key investment criteria. The investment group began a thorough analysis of the market and the properties eligible for historic tax credits and found The Spinning Mill, a once prominent mill building that had fallen into disrepair. After continued analysis of the city and the site itself, the investment group entered into contract to purchase the site and convert the building into 32 residential apartments with an intended completion of Spring 2024.

The Skowhegan mill acquisition is a key use case for *HTC View* as a way for investors to quickly analyze and track thousands of market simultaneously and funnel down the markets that most fit their specific investment criteria.

*Figure 5.1: Current condition of 7 Island Avenue, Skowhegan Maine. Photo May 2022*


*Figure 5.2: Computer generated 3D rendering of 7 Island Avenue, Skowhegan Maine, 04976 intended to be complete in Fall 2024*

## 5.3 PLATFORM LIMITATIONS

*HTC View* program illustrates how web data scraping can be effectively used to advance an investment and acquisitions strategy. The tool enables a user to conduct an analysis on hundreds of markets simultaneously, which would normally take hundreds of hours to complete. However, there are several technical limitations due to the unspecified nature of some of the National Parks Service underlying dataset. The way in which historic properties are logged and saved in the NPS database leaves out key pieces of information that would enhance the performance of the *HTC View* platform. One key piece of missing data is that the buildings that are listed within a historical district are not easily discernable or well organized. Most of the archival data on the NPS website, stems from physical photocopies or pictures of the original historic application accepted by the National Parks Service whenever the historic building or district was added to the registry. Further, these historic districts generally do not contain accurate latitude and longitude boundaries that would make mapping and tracking properties that fell within them easier. Instead, the boundaries for these specific districts are denoted

through non-uniform written descriptions of the streets and areas that do and do not fall within them.

While these underlying data issues ultimately limit the reliability of *HTC View* there are a number of workarounds that ensure that the user is working with a reliable data set. Building these new 'data-cleaning' tools and functions will be a key part of *HTC View's* product roadmap and future development. The opportunity to advance this tool as well if and when the National Parks Service improves their cataloging system would also improve the platform's performance significantly.

# CHAPTER 6: CONCLUSION & RECOMMENDATIONS

## 6.1 CONCLUSION

The goal of this research is to examine the effectiveness of web data scraping tools when applied to core functions of real estate investment professionals. To determine this, the author has taken on the development of a web scraping tool focused on real estate acquisitions that qualify for historic tax credits through *HTC View.* Through this highly specific avenue, it is clear that there is significant value in data web scraping applications and the technology should be adopted on a wider scale. This is particularly true of highly repeatable, data-rich, applications like real estate acquisitions in which many geographies and opportunities are being compared simultaneously. *HTC View* focused exclusively on the use of big data for real estate acquisitions, however it seems evident that the same tools could be applied to other areas of the industry (i.e. asset management) and prove equally valuable.

The effectiveness of these web scraping tools not only rely on the richness and quality of the underlying data but on a clear specifically defined output. Determining the key variables one needs for effective web scraping analysis, and then determining the best sources to reliably extract this data, is perhaps the most challenging aspect of building an effective web scraping algorithm.

Future development on *HTC View* will focus on adding additional key data sources, and enhancing the user experience to make new acquisition discovery more effective. Through this research, it is evident that there is significant value

in web scraping data applications, for real estate purposes and beyond, and should become a fundamental part of data analysis and processes as a core means of automated time optimization.

## REFERENCES

Kitchen, R, 2014, The real time city? Big data and smart urbanism, GeoJournal, vol. 79, no. 1, pp. 1-14.

THE ECONOMIST, May 2017 . The world's most valuable resource is no longer oil, but data. *The Economist Magazine,* May 2017.

Munzert S, Rubba C, Meissner P, Nyhuis D. 2014. Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons

MOHANRAM, P. S. 2020. A Brave New World: The Use of Non-traditional Information in Capital Markets. *World Scientific Book Chapters***,** 217-237.

PATEL, J. M. 2020. Introduction to Web Scraping. *Getting Structured Data from the Internet.* Springer.

SNELL, J 2013. Use of Online Data in the Big Data Era: Legal Issues Raised By the Use of Web Crawling and Scraping Tools for Analytics Purposes

WINSON-GEIDEMAN, K., KRAUSE, A., LIPSCOMB, C. A. & EVANGELOPOULOS, N. 2017. *Real estate analysis in the information age: techniques for big data and statistical modeling, Routledge.*

Zembowicz, R., and Zytkow, J. 1996. From Contingency Tables to Various Forms of Knowledge in Databases. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 329–351. Menlo Park, Calif.: AAAI Press.