Smart Search for Rental Properties

Zillow

Mahendra Chagam

Executive Summary

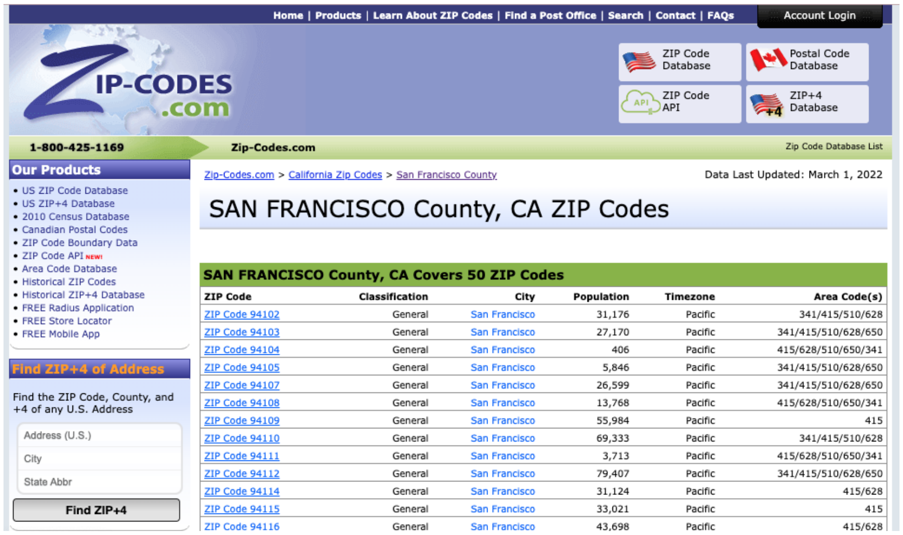
Web scraping, the process of using computer programs to automate data extraction from websites, can be very useful for a variety of users having different needs. Real estate professionals, renters and also landlords can benefit from the dataset we built as the housing/rent market is one of the most dynamic ones. In our project, we have utilized diverse web scraping techniques such as MongoDB, json, BeautifulSoup and API. For this report, we have limited the rent listings to San Francisco, California in the United States. We have scraped **1) a Zip code repository website 2) Zillow website 3) Google Maps API**. Combining all of the information collected from the fore-mentioned websites and API, we aim to provide a user looking for properties to rent with the nearest and most relevant set of properties along with the price, distance and various other informative features that help him/her shortlist the properties they want to visit and check out or even make a buy/rent decision. That is, we aim to mainly benefit zillow users or anyone who is looking for properties to buy or rent by helping them make more informed decisions.

Background, Context, and Domain Knowledge

We will look for rental properties on Zillow, one of the leading home rental and real estate websites in the market. Scraping rent listings on Zillow.com enables us to search properties in any location even though in the report, we have limited the location to where we live, San Francisco, CA. We are able to extract detailed information such as full addresses, price, number of bedrooms and bathrooms, and all other information available on the website. In this project, we have scraped all of San Francisco zip codes and the listings within those zip codes. A common use case of this project would be when a user wants to see the nearest rental listings by providing their current address or the neighborhood they want to live in. The listings can also be filtered by the distance specified by a user from a certain address or zip code. Users will be able to see and compare different features and characteristics of the listings such as price or number of bedrooms and bathrooms. The database we have created will provide the users with a low to no cost tool, not to mention the fast speed of the web scraping.

Data Sources, Web-scraping and Processing Details

To begin with, as we have to capture all the rental listings present in the entire San Francisco, we need to collect zip codes. To obtain them, we used the zipcodes.com website (https://www.zip-codes.com/county/ca-san-francisco.asp). This website provides all the zipcodes with respect to any county in the entire United States. We are limiting the city to San Francisco for this project; however, by changing the county part in this link, we can easily expand the scope of our project and produce results for a different city or multiple cities as well should a reader of this report would like to apply this web scraping model to other cities.



Zip-Codes.com > California Zip Codes > San Francisco County

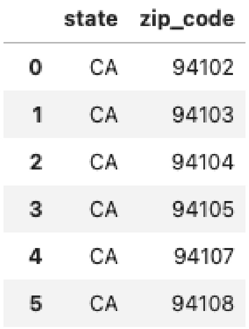
Data Last Updated: March 1, 2022

SAN FRANCISCO County, CA ZIP Codes

SAN FRANCISCO County, CA Covers 50 ZIP Codes

ZIP Code	Classification	City	Population	Timezone	Area Code(s)
ZIP Code 94102	General	San Francisco	31,176	Pacific	341/415/510/628
ZIP Code 94103	General	San Francisco	27,170	Pacific	341/415/510/628/650
ZIP Code 94104	General	San Francisco	406	Pacific	415/628/510/650/341
ZIP Code 94105	General	San Francisco	5,846	Pacific	341/415/510/628/650
ZIP Code 94107	General	San Francisco	26,599	Pacific	341/415/510/628/650
ZIP Code 94108	General	San Francisco	13,768	Pacific	415/628/510/650/341
ZIP Code 94109	General	San Francisco	55,984	Pacific	415
ZIP Code 94110	General	San Francisco	69,333	Pacific	341/415/510/628
ZIP Code 94111	General	San Francisco	3,713	Pacific	415/628/510/650/341
ZIP Code 94112	General	San Francisco	79,407	Pacific	341/415/510/628/650
ZIP Code 94114	General	San Francisco	31,124	Pacific	415/628
ZIP Code 94115	General	San Francisco	33,021	Pacific	415
ZIP Code 94116	General	San Francisco	43,698	Pacific	415/628

We found the table structure on the web page above and went through each table detail to capture the zip codes of San Francisco. We used BeautifulSoup to convert the captured web page structure into a html format to parse through them. After saving the page as a beautifulsoup object, we utilized it to find zip codes and stored them in our dataframe and the output is as below(showing the top 6 rows).



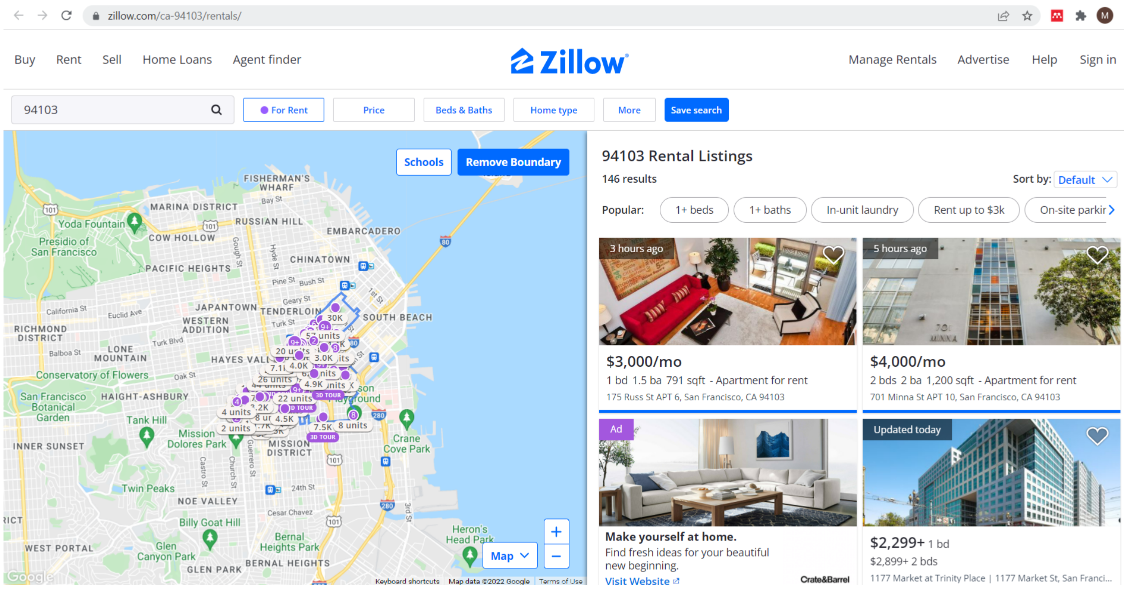
	state	zip_code
0	CA	94102
1	CA	94103
2	CA	94104
3	CA	94105
4	CA	94107
5	CA	94108

Then, we leveraged zip codes captured above for San Francisco to find the rental listings on Zillow.com. We created a generic URL to iterate through all the zip codes in San Francisco.

URL Structure : *https://www.zillow.com/"State"-"Zipcode"/rentals/*

For example, for California with zip code 94103, the link would look like:

https://www.zillow.com/ca-94103/rentals/. Below is the screenshot of the search result.



94103 Rental Listings

146 results

Sort by: Default

Popular: 1+ beds | 1+ baths | In-unit laundry | Rent up to $3k | On-site parkir

3 hours ago

$3,000/mo

1 bd 1.5 ba 791 sqft - Apartment for rent

175 Russ St APT 6, San Francisco, CA 94103

5 hours ago

$4,000/mo

2 bds 2 ba 1,200 sqft - Apartment for rent

701 Minna St APT 10, San Francisco, CA 94103

Ad

Make yourself at home.

Find fresh ideas for your beautiful new beginning.

Visit Website

Crate&Barrel

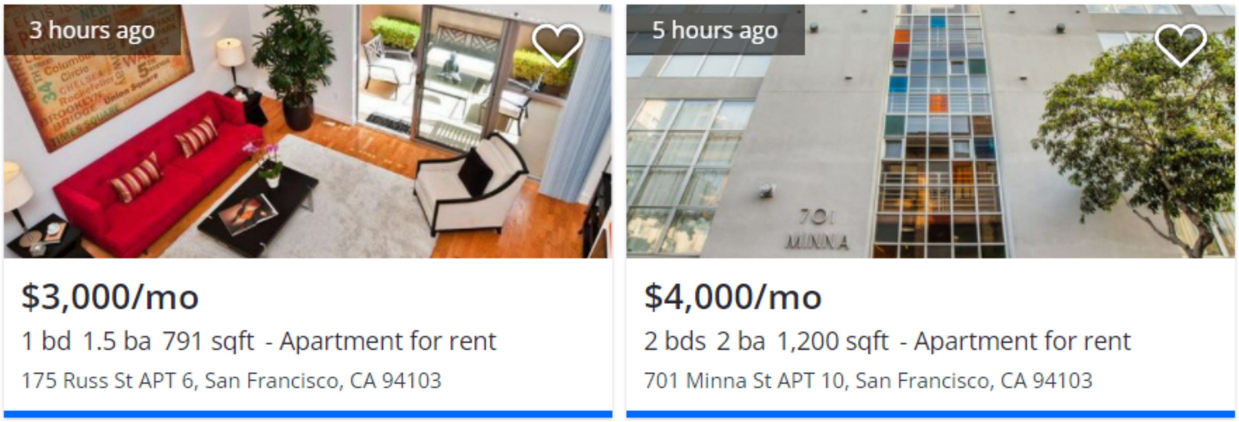
Updated today

$2,299+ 1 bd

$2,899+ 2 bds

1177 Market at Trinity Place | 1177 Market St, San Franci...

Once we reach the search result page for a specific zip code, we see the total number of listings and the actual listings on the right side. Our objective is to capture the information from all of the listings in the search result. We tried scrapping the information but the challenge was that we could capture only the top 10 listings as zillow has the feature of loading only when you scroll. Therefore, we had to find a way around to get the details in the page. After examining the page structure, we found out that there is a json object embedded inside the page to retrieve information while scrolling. We then captured the json object and parsed through that to get the listing information.



3 hours ago

$3,000/mo

1 bd 1.5 ba 791 sqft - Apartment for rent

175 Russ St APT 6, San Francisco, CA 94103

5 hours ago

$4,000/mo

2 bds 2 ba 1,200 sqft - Apartment for rent

701 Minna St APT 10, San Francisco, CA 94103

For every listing we captured :

Zpid, imgSrc, detailUrl, statusType, statusText, addressStreet, zipcode, city, state, latitude, longitude, price, bathrooms, bedrooms, livingArea, homeType, homeStatus, daysOnZillow isFeatured, shouldHighlight, isRentalWithBasePrice, isPremierBuilder, currency, and country.

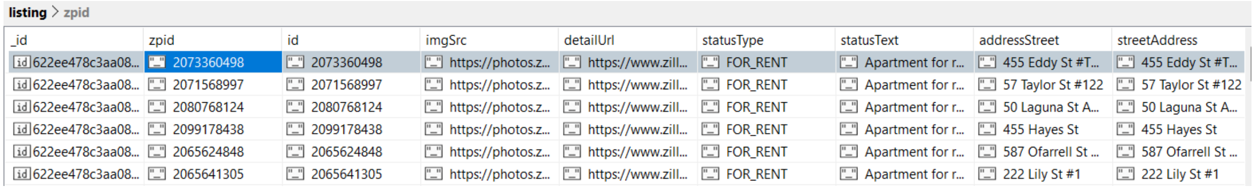
Once we capture the details from the page, the next step is to do pagination to capture all the listings from multiple pages. We used **a pagination variable** in the generic URL to iterate through multiple pages. The URL looks like:

Pagination URL : https://www.zillow.com/"State"-"Zipcode"/rentals/"pagenumber"_p

We kept on increasing the page number until the captured listings are equal to the total number of listings of the given zip code.

After capturing the data, we used Mongo DB to save the data. We used NoSql document model to store data into the database. We choose NoSql database as it can handle a large amount of data at high speed. MongoDB stores the data which has a hash table encoded with it, thereby enabling faster retrieval of data. The other database models are relational, hierarchical, network, object oriented and entity relationship. Out of all these relational databases are the most popular which need proper structure and relationships built across the tables to retrieve the information. The relational databases are ideal for performance; however, they require a lot of maintenance as the data keeps growing whereas the NoSql MongoDB database needs minimal maintenance.

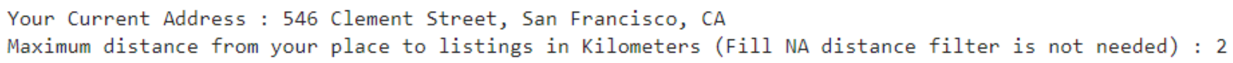
In MongoDB, we created a database zillow and also a collection named listing. Below is the view of the table from studio3T.



listing > zpid

_id	zpid	id	imgSrc	detailUrl	statusType	statusText	addressStreet	streetAddress
622ee478c3aa08...	2073360498	2073360498	https://photos.z...	https://www.zill...	FOR_RENT	Apartment for r...	455 Eddy St #T...	455 Eddy St #T...
622ee478c3aa08...	2071568997	2071568997	https://photos.z...	https://www.zill...	FOR_RENT	Apartment for r...	57 Taylor St #122	57 Taylor St #122
622ee478c3aa08...	2080768124	2080768124	https://photos.z...	https://www.zill...	FOR_RENT	Apartment for r...	50 Laguna St A...	50 Laguna St A...
622ee478c3aa08...	2099178438	2099178438	https://photos.z...	https://www.zill...	FOR_RENT	Apartment for r...	455 Hayes St	455 Hayes St
622ee478c3aa08...	2065624848	2065624848	https://photos.z...	https://www.zill...	FOR_RENT	Apartment for r...	587 Ofarrell St ...	587 Ofarrell St ...
622ee478c3aa08...	2065641305	2065641305	https://photos.z...	https://www.zill...	FOR_RENT	Apartment for r...	222 Lily St #1	222 Lily St #1

Once we created the database, we made a customer facing search portal where a customer comes to the page and gives his/her address and the range in which he/she is searching for listings from their location. Below is the example:



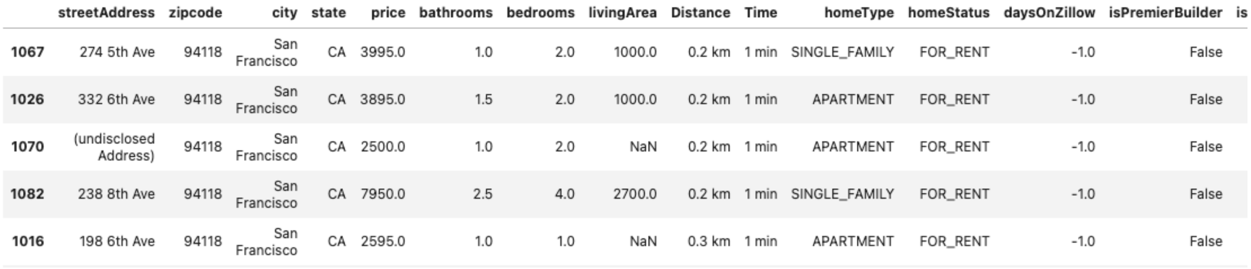
```
Your Current Address : 546 Clement Street, San Francisco, CA
Maximum distance from your place to listings in Kilometers (Fill NA distance filter is not needed) : 2
```

We received details from the customer, queried our database for all the listings and found the distance from the customer's location to all the listings by using google maps API.

API structure : https://maps.googleapis.com/maps/api/distancematrix/json?

destinations="customer_address"&origins= "listing_address"&key="apikey"

Then, we captured the distance and time to travel from the customer's location to all the listings from API, added that information to the database query result and filtered the result for the distance range given by the customer. Finally, we displayed the result sorted from the nearest to farthest from the customer's location. Below is the display for the address input and the distance range given by the customer, sorted by distance .



	streetAddress	zipcode	city	state	price	bathrooms	bedrooms	livingArea	Distance	Time	homeType	homeStatus	daysOnZillow	isPremierBuilder	is
1067	274 5th Ave	94118	San Francisco	CA	3995.0	1.0	2.0	1000.0	0.2 km	1 min	SINGLE_FAMILY	FOR_RENT	-1.0	False	
1026	332 6th Ave	94118	San Francisco	CA	3895.0	1.5	2.0	1000.0	0.2 km	1 min	APARTMENT	FOR_RENT	-1.0	False	
1070	(undisclosed Address)	94118	San Francisco	CA	2500.0	1.0	2.0	NaN	0.2 km	1 min	APARTMENT	FOR_RENT	-1.0	False	
1082	238 8th Ave	94118	San Francisco	CA	7950.0	2.5	4.0	2700.0	0.2 km	1 min	SINGLE_FAMILY	FOR_RENT	-1.0	False	
1016	198 6th Ave	94118	San Francisco	CA	2595.0	1.0	1.0	NaN	0.3 km	1 min	APARTMENT	FOR_RENT	-1.0	False	

For the extent of this project we focused on listings of San Francisco. However, this concept can be scaled to any city in the United States. In addition, we have provided only a few search filters. We can include a filter for price range, square footage, apartment type, bed rooms, and bath rooms. This allows users to find the exact type of listing they are looking for.

Dataset Application and Business Value

Zillow has a huge presence in the real estate industry with over 100 million properties in its database. Whether it is to buy or rent a house, it is an extremely important decision as it directly affects our day-to-day lives and literally one’s “home”. For zillow users, it is important that they are making smart and informed decisions by ensuring that they are understanding how the real estate/rental market is changing. Our project will help individuals make better investment decisions as cost associated with property or rental usually takes up a big chunk of the paycheck. It will also benefit real estate agencies or brokers, this database implementation could help them track the listings and identify trends in the market.

Not only that, we can use this dataset to predict the price or any other noticeable trends of the properties based on zip codes. This information can essentially help users to prevent themselves from paying over priced costs for buying or renting a place than market standards. Also, from the opposite side of the relationship, those who are looking to sell or rent out their property can also obtain a better idea of the market standards and accordingly determine their listing prices or at least have an idea where their property stands compared to other listings available on the market. In this sense, they will have a better chance of selling the property without wasting time and energy trying to meet the desired standards of the market.

Apart from the business use cases here are a few questions that help with the efficient usage of dataset captured:

1. What is the average rental price in any given zip code?
2. What are the types of houses available in San Francisco or any zip code?
3. How is the living area related to rental price?
4. How is the number of bedrooms and bathrooms related to the rental price?

Conclusion

We have scraped 3 different sources: 1) Zip code repository website 2) Zillow website 3) Google Maps API to achieve automatic data extraction of rental listings. This dataset will benefit individuals who are looking to rent or purchase a property and save their time and energy by offering detailed and important information from numerous listings from Zillow. We went beyond simply compiling information from the websites and provided a more important, customized and most relevant set of properties along with the distance, price and other numerous informative features. There were some challenges in the process of building the web-scraping dataset; for instance, having to scroll down the page to be able to see all of the listings on a page. Despite the challenges, we built a successful database by incorporating 3 different data sources and gathering all of the information for a certain zip code or in our report, for zip codes of San Francisco. Given the importance of choosing the right home/property for any individual, our database will have a greater impact by helping users make an informed decision.