STA 220 Final Project: Getting data to assist in Apartment Hunting

Kevin Su ID#914935300

March 21, 2023

Background

I'm expecting to graduate from the MS Statistics program in June of 2023, and I have been lucky enough to secure an offer of employment from the National Agricultural Statistics Service with the US Department of Agriculture, the office I have been interning at for the past year or so in Sacramento, California. One of their agency policies is that unless there's a severe shortage of personnel at one of their offices, onboarding employees are placed somewhere other than the office where they intern at. I have been assigned to the Olympia, Washington, regional office. As such, I will need to find a suitable apartment for me and my girlfriend, who is moving up with me, to live in.

Objective

The objective of this project is to come up with an effective way to aggregate and filter for desirable characteristics of apartments on the market. There already exist quite a few websites for this, but I'm going to focus on two in particular: Apartments.com and Zillow. This is because Apartments.com allows for many different potential filters that Zillow doesn't consider while at the same time listing an easily accessible phone number. Zillow, unlike Apartments.com, breaks down the cost of units from the same apartment complex depending on the number of bedrooms in the unit and also offers relatively retrievable coordinates of a location that I may be interested in plotting on a map to consider the driving distance from my workplace. In general, I would like a concise output of various places that I would consider living in using Apartments.com's filter while being able to integrate some model-level information using Zillow.

Living Considerations

Before searching for any apartments, it is important to determine what the filters I would like to consider are. We could consider a one-bedroom apartment, but we have many items that we are bringing up and would like some extra storage space. Because of the need for more space and potentially having to set up a remote office, a second bedroom would probably be ideal. Additionally, having a second bedroom would allow us to store and hide away any household objects that are potentially dangerous to our two cats. Since we are bringing our two cats up, we would need wherever we choose to be welcoming of cats as pets. Because of the small size of the city, we would consider living anywhere within the Olympia-Lacey-Tumwater metro area, as long as we can have somewhere to park our cars. Additionally, my girlfriend demands that wherever we live, we must have access to a community pool. Because I'm being offered an initial starting salary of approximately $63,526 for the first year, subject to an inflation adjustment, I believe my ideal cost for rent is limited to roughly $1600/month. However, because of my lack of debts and some potential contribution from my girlfriend, I think $2100 is the absolute upper limit of what we can afford. To sum it up, we need to consider any cat-friendly apartment complex that offers one or two-bedroom units for under $2100, has on-site parking, and has access to a pool.

Retrieving Data from Apartments.com

The Apartments.com search function makes an API call requesting a POST method where it expects that our payload that I provide has some coordinate search boundaries, minimum number of bedrooms, maximum number of bedrooms, pet-friendly status, and an int value that corresponds to the various combinations of amenities we expect. The response initially appears to be in JSON format and returns a dict with around 16 keys. However, the key that contains all of the relevant data I am interested in is the 'PlacardState' key, which itself when called returns a dict where all of the relevant data is actually in a key called 'HTML', which itself is a string that contains all of the data in HTML format. To

process the data into a useable format, I used BeautifulSoup and find_all to find the names of apartment complexes which were within 'span's with class "js-placardTitle title", addresses of apartment complexes which were in 'div's with class "property-address js-url", price ranges which had 'p' tags with class "property-pricing", phone numbers which had 'a' tags with classes similar to but not exactly matching "phone-link". Because phone numbers had multiple iterations in different formats, the use of regex was necessary to compile all phone numbers. Additionally, addresses were decomposed into their Street, City, State, and ZIP components. The resultant lists were converted into pandas Series objects and concatenated into a pandas DataFrame, a truncated example of which is included below.

]:

	Name	Address	City	State	ZIP	Apartments-Price	Phone
0	Smyth	3425 Polo Club Ln SE	Olympia	WA	98501	$1,730	(360) 515-3840
1	Breckenridge Apartments	2820 Tuscany Ln SW	Tumwater	WA	98512	$1,470 - 1,800	(360) 515-4235
2	Imber at Union Mills	8519 Oya Ln SE	Olympia	WA	98513	$1,577 - 2,385	(360) 634-0417
3	Montair at Somerset Hill	1704 Barnes Blvd SW	Tumwater	WA	98512	$1,423 - 2,401	(360) 634-0999
4	Callen	1404 Brittany Ln NE	Lacey	WA	98516	$1,700 - 1,815	(360) 995-1232
5	Lacey Park Apartments	5001 College St SE	Lacey	WA	98503	$1,400 - 1,500	(360) 515-3837
6	Toscana Apartments I	6979 Birdseye Ave NE	Lacey	WA	98516	$1,595 - 2,195	(360) 995-1079
7	Abbey Rowe Apartments	9320 Windsor Ln NE	Olympia	WA	98516	$1,831 - 1,974	(360) 634-4503
8	Gayteway at Hawks Prairie	8825 Martin Way E	Lacey	WA	98516	$1,599 - 1,905	(360) 634-4758
9	Britton Place	6655 Britton Pky NE	Lacey	WA	98516	$1,862 - 5,681	(360) 634-0968
10	The Dakota	6205 Pacific Ave SE	Lacey	WA	98503	$1,660 - 2,167	(360) 972-3564

Apartments.com extract of Apartment complex listings meeting filter criterion

Retrieving Data from Zillow

The next objective is to retrieve Zillow's data for the same region, especially the price for each unit class and coordinates. Here, there is a major problem with getting the data from Zillow. Scraping the page directly will give us the prices for units by unit class but will not give us coordinate information. Querying their search API leads us to a series of redirects and multiple protections against automated querying, and we receive no data despite returning an HTTP Status code of 200, normally indicating success. Notably however, their Request URL alone directly gives us what we want: an output of the data

we want in JSON format directly. However, it cannot be easily accessed through the requests function, so we save the JSON text output as a JSON file to be called.

When processing the Zillow data, it is relatively straightforward to extract the Address, Latitude, and Longitude of the returned apartment records. However, we have to create duplicates of the records according to the number of different classes of units they sell (i.e. a complex that offers both 1 and 2 beds would have to have two separate entries). We flatten the dict that contains the unit class data and restructure the data as such. A truncated example of the pandas DataFrame containing the extracted information from Zillow is displayed below. Note that the entries include Studio (marked as 0 bed) and three-bedroom units. This will later be processed to remove the Studio apartments after merging. As we are not looking to Zillow to filter data, only as a source of supplemental data to the Apartments.com set, some entries in the Zillow set are outside of our ideal filters in order to not accidentally exclude potential overlapping data. These extra entries will later not make it into the final set in question, except for possible three-bedroom units as a brief cost exercise rather than serious consideration.

	Name	Address	State	City	Zillow-Price	Beds	Latitude	Longitude
0	Imber at Union Mills	8519 Oya Ln	WA	Olympia	$1,577+	1	46.994087	-122.734215
0	Imber at Union Mills	8519 Oya Ln	WA	Olympia	$1,869+	2	46.994087	-122.734215
1	Martingale Apartments	8675 Litt Dr SE	WA	Olympia	$1,690+	0	47.056120	-122.756280
1	Martingale Apartments	8675 Litt Dr SE	WA	Olympia	$1,795+	1	47.056120	-122.756280
1	Martingale Apartments	8675 Litt Dr SE	WA	Olympia	$2,000+	2	47.056120	-122.756280
2	Switchback Apartments	7127 32nd Ave NE	WA	Olympia	$1,650+	1	47.076523	-122.786500
2	Switchback Apartments	7127 32nd Ave NE	WA	Olympia	$2,025+	2	47.076523	-122.786500
3	Toscana Apartments I	6979 Birdseye Ave NE	WA	Olympia	$1,660+	0	47.070170	-122.787860
3	Toscana Apartments I	6979 Birdseye Ave NE	WA	Olympia	$1,595+	1	47.070170	-122.787860
3	Toscana Apartments I	6979 Birdseye Ave NE	WA	Olympia	$2,075+	2	47.070170	-122.787860
4	Chambers Reserve Townhomes	3725 Wildspitz St SE	WA	Lacey	$2,595+	3	47.013206	-122.831590

Zillow listings extract

Merging and other Processing

A problem that I encountered was spelling discrepancies between apartment names and addresses on Zillow and apartment names and addresses on Apartment.com, as well as some artificially

injected space characters. To try to minimize this discrepancy, each double and triple space was reduced to one space. Additionally, the Zillow-quoted prices of the units were reformatted via regex to only have numeric values in order to convert the column to float type for some further investigation. Subsequently, the Apartments.com and Zillow pandas DataFrames were merged twice, once on matching addresses, and another time on matching apartment complex names. These entries were stacked and duplicate entries as well as entries with zero bedrooms were deleted. This method of trying to ensure no entries were lost due to double failures in data entry is not foolproof, and perhaps storing it in some database format to use a SQL "LIKE" operator might have been more appropriate. However, even this would not be foolproof, as we will later see in post-review caveats.

Following this, the original Apartments.com DataFrame was merged with this merged table in order to not lose any entries from the original Apartments.com search. These entries that Zillow did not pick up shouldn't be dropped as I might still be interested in these-I just wouldn't be able to plot them on a map because I lack the coordinate information, and the few that weren't picked may warrant further manual investigation if our initial search turns up fruitless. If it were not for the fact that maps and geo-services APIs are not free to use (or are freemium at best), I would be more inclined to hunt down individual coordinates for these locations as well, given that the address data is present.

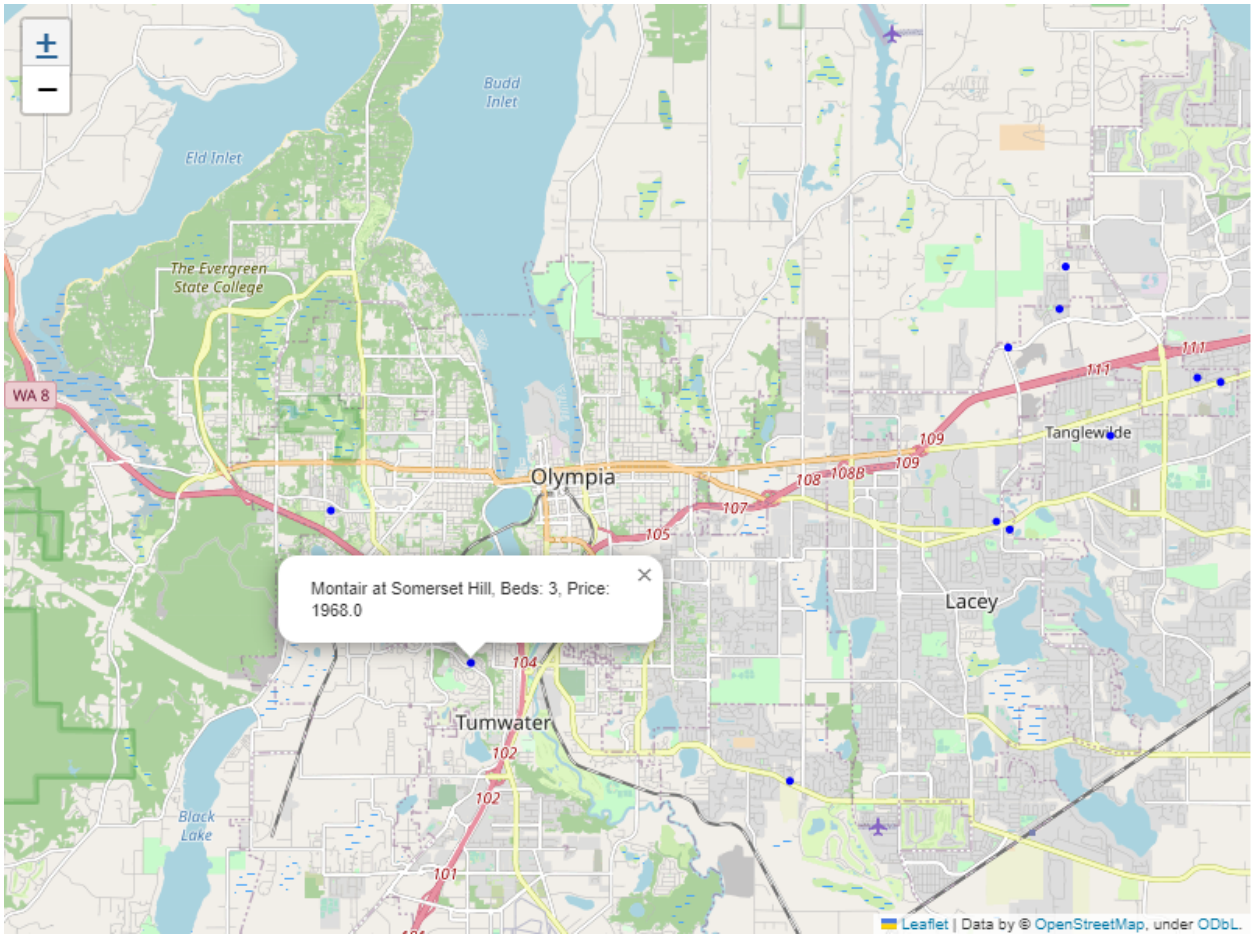
A truncated example of the final merged pandas DataFrame is shown on the following page. This final DataFrame has 38 entries of different unit classes offered by 25 different complexes. However, one complex that offers one unit class is a "55+ community" and would not be considered. Additionally, there is no Zillow price information on this, so it would not factor into further investigation on the choice of number of bedrooms.

	Name	Address	City	State	ZIP	Beds	Apartments-Price	Zillow-Price	Latitude	Longitude	Phone
0	Smyth	3425 Polo Club Ln SE	Olympia	WA	98501	2	$1,730	1726.0	46.999260	-122.847270	(360) 515-3840
1	Smyth	3425 Polo Club Ln SE	Olympia	WA	98501	3	$1,730	2129.0	46.999260	-122.847270	(360) 515-3840
2	Breckenridge Apartments	2820 Tuscany Ln SW	Tumwater	WA	98512	NaN	$1,470 - 1,800	NaN	NaN	NaN	(360) 515-4235
3	Imber at Union Mills	8519 Oya Ln SE	Olympia	WA	98513	1	$1,577 - 2,385	1577.0	46.994087	-122.734215	(360) 634-0417
4	Imber at Union Mills	8519 Oya Ln SE	Olympia	WA	98513	2	$1,577 - 2,385	1869.0	46.994087	-122.734215	(360) 634-0417
5	Montair at Somerset Hill	1704 Barnes Blvd SW	Tumwater	WA	98512	1	$1,423 - 2,401	1423.0	47.017014	-122.917534	(360) 634-0999
6	Montair at Somerset Hill	1704 Barnes Blvd SW	Tumwater	WA	98512	2	$1,423 - 2,401	1582.0	47.017014	-122.917534	(360) 634-0999
7	Montair at Somerset Hill	1704 Barnes Blvd SW	Tumwater	WA	98512	3	$1,423 - 2,401	1968.0	47.017014	-122.917534	(360) 634-0999
8	Callen	1404 Brittany Ln NE	Lacey	WA	98516	1	$1,700 - 1,815	1763.0	47.059850	-122.757470	(360) 995-1232
9	Callen	1404 Brittany Ln NE	Lacey	WA	98516	2	$1,700 - 1,815	1837.0	47.059850	-122.757470	(360) 995-1232
10	Lacey Park Apartments	5001 College St SE	Lacey	WA	98503	NaN	$1,400 - 1,500	NaN	NaN	NaN	(360) 515-3837

Truncated merged dataframe of all complexes that fit the criterion.

Mapping

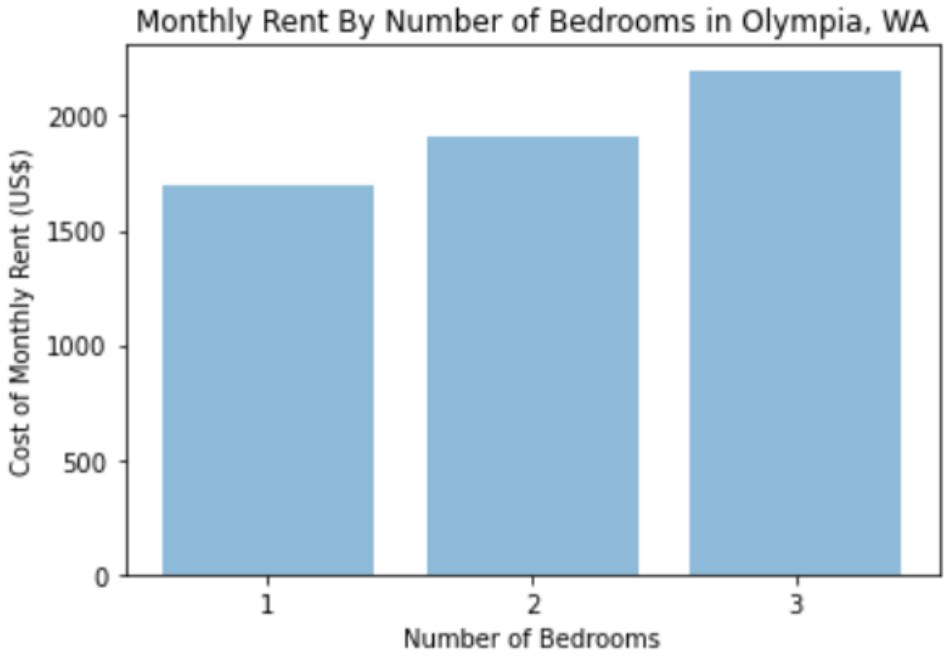
From this, we can now plot these locations on a map and consider the possible distance on my commute. This could factor into our decision on where to live, but ultimately, Olympia is a small enough city that commute times are not too brutal in most circumstances. We plot the locations through the use of the folium library. A *picture* is shown on the next page, but **a working HTML version will also be separately attached with this submission and available on github at https://github.com/Kvnsu/STA-220-Final-Project-Winter-2023/tree/main where it will need to be downloaded and separately opened**. **The file is called "OlympiaLocs.html."** From the map, it appears that "Montair at Somerset Hill" is the closest to my workplace (set as the center of the image, unmarked). They offer one, two, and three bedroom options at prices of 1423, 1582, and 1968 per month, respectively. This is where we'll probably start our search, though Woodland, the dot on the left, may also be worth considering as it has both one and two-bed units for 1629 and 1707 respectively.



Further Considerations

We are also considering having some extended family stay with us for a while due to family dynamics, care, and access to next of kin. To do this, we evaluate roughly how much each increase in a bedroom costs simply by taking the grouped mean of the Zillow-price column on the number of bedrooms. On average, a 1-bedroom apartment that fits our criteria costs $1700/month to rent, while a 2-bedroom costs $1907/month, and a 3-bedroom costs $2200/month. Although we've filtered specifically for one or two-bed apartments that cost less than $2100 when doing our search, which will make these numbers not quite true overall, the important point is that these 3-bed units are offered in the same complexes that offer the two-bed and 1-bed units, which can serve to demonstrate that the

third bedroom might cost meaningfully more by margin than the second (300 for the third bedroom as opposed to 200 for the second), as shown in the matplotlib bar plot below.



Caveats

The subject and method of our search is not foolproof and some intriguing locations that are worth consideration may fall through the cracks. For example, Breckenridge Apartments, a complex that was listed on the Apartments.com website, did not show up on the Zillow extract. This is because the complex name was not entered at all in Zillow, and the addresses would not show up as a match when merging because the addresses are given a new street number per unit, rather than an individual apartment number because they exist as a townhouse-styled complex despite being marketed as apartments. There's not a clear way using these methods that I can easily conceive to resolve this problem without introducing other sources for error.

Additionally, the number of sources could be increased as Apartments.com and Zillow are not necessarily all-encompassing of the number of potential apartment listings out there. Despite this, this

piece of code and project helps provide me with a solid base on which to pursue further places to investigate.

Extras that are not quite relevant to this/Comments

Also, I mentioned having cats earlier, and I feel that it is probably a violation of international law to mention cats without showing pictures, so here they are. The gray one on the left is named Scrambles and the one on the right is named Cheddar. Cheddar likes to jump and is generally scared of everything. Scrambles loves food and is actually very fat even if he doesn't look that way in the photo. He's scared of nothing in this world… except Cheddar. I figure I've hit my 8 pages (This is page 9) and so it wouldn't be inappropriate to put them here as I couldn't find a way to work them into the body of the paper.

Apologies for using Google Docs. I made a concerted attempt to try to get this to work in Jupyter Notebook as well as R Markdown but the indentation not copying over well to Rmd for use with reticulate made it impractical, and I wrecked my Anaconda environments trying to get R to work with Jupyter Notebook for Homework 4, so exporting to PDF no longer worked on Jupyter.