Ann Arbor Housing Survey: Final Report

"The Rent is Too Damn High" — Ann Arbor edition

Michigan Data Science Team, Winter 2020

[link to blog post] [link to article] [link to repository]

Project Lead: Eric Chen

Team Members: Ethan Burt, Iris Derry, Tejas Kulkarni, Anthony Ng, Mariah Zeweke, Lily Zhai

Contributors: Grant Barry, Manav Bhatia, Dylan Kalten, Saman Verma, Darryl Wong, Vivian Wong

TABLE OF CONTENTS.

1. INTRODUCTION.

Welcome to Ann Arbor, Michigan: home of the Wolverines, food lover's paradise, hotbed of political activism, and … really expensive rent?

Any student at the University of Michigan knows how difficult it is to find affordable housing. While residence halls can be a convenient option, they can also be pricey: even the cheapest undergrad housing option, a triple, comes out to about $9910, or about $1200 per month.

Looking off-campus can be cheaper, but that poses its own challenges. Off-campus housing is often located in inconvenient locations, while locations closer to campus can be quite expensive — don't even get me started on some of the bougie high-rises that have been popping up everywhere! 🤯

With this in mind, our team set out to conduct a detailed analysis of the state of off-campus housing. Questions we sought to answer included:

- How expensive is it to actually live off-campus? Is it really as bad as it's made out to be? (aka "is this real life, or is this just fantasy?")
- What interesting trends can we observe when we sort or filter by different features?
- What features seem to influence price the most? Can we predict the price based on these features?
- It's no secret that realty websites are trying to sell you something! To this end, what interesting insights can we extract from the posted descriptions of each listing?

2. METHODOLOGY AND DATA COLLECTION.

The data consists of about 9000 property listings scraped from aggregator sites such as Craigslist, ShowMeTheRent, apartments.com, and UMich's own off-campus housing website. We used BeautifulSoup, a smattering of regex, and some good old-fashioned elbow grease to extract the data, which were in various forms of completeness; some of the Craigslist listings were missing addresses or prices entirely! Many of the features were quite sparse, and ended up being dropped at some point in the project.

Our features included price, address, number of bedrooms, rental company if applicable, etc. A full list of features can be found in the `README` in the linked repository, under the section "A Look at the Data".

Later on in the project, we geocoded each listing's address so that we could calculate additional features from the coordinates, such as distance to various landmarks; these include the Michigan Union, the Central Campus Recreation Building (CCRB), and the Shapiro Undergraduate Library (UgLi). The full list of landmarks can be found in the repository under `geocoding/distance-neighborhood.ipynb`. This did require us to drop any entries that were missing or had incomplete addresses.

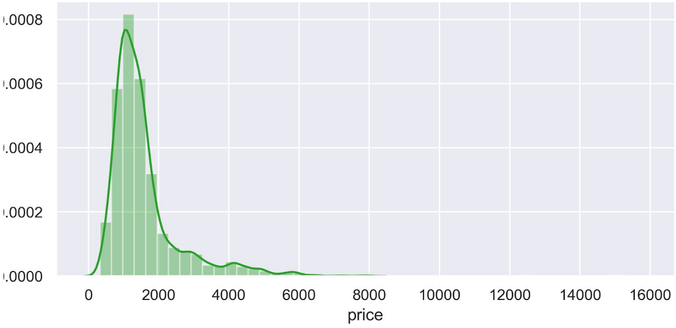
Throughout this report, assumptions and technical remarks will be in dark gray.

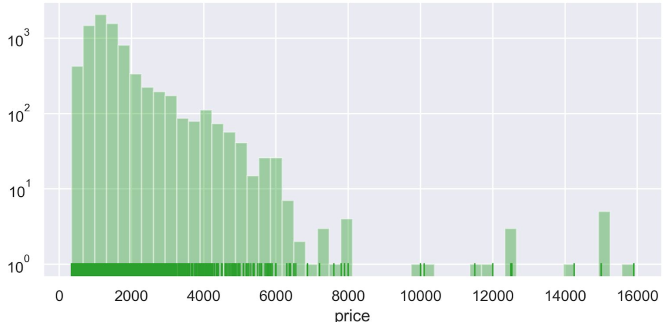
3. FINDINGS.

How expensive is Ann Arbor, really?

The most obvious thing to do was look at the distribution of prices in our data. The bottom graph is plotted on a log-linear scale in order to highlight the tail of the distribution. Overall, a vast majority of the listed prices were between $650 and $2000.

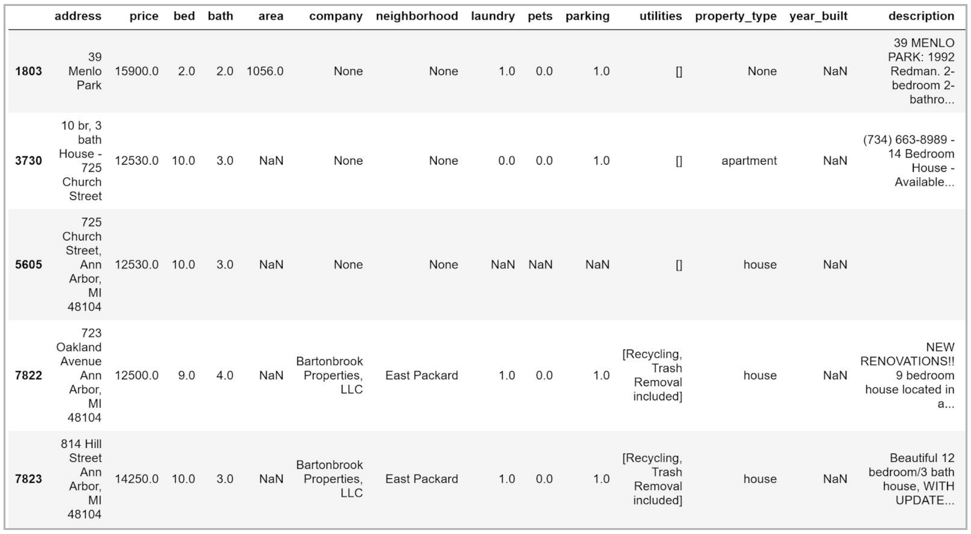
There were some strange entries in the data. From anecdotal experience, not many students pay under $500 per month for housing, so we (somewhat arbitrarily) chose $300 as our lower bound and excluded the remaining 20 entries, many with listed prices of 0 or 1. These strange values are almost definitely due to error in the data collection process.





We had some crazy outliers in the data; we'd never heard of anyone paying ~$15000 for housing! Upon further examination, most of these listings turned out to be large houses with 9 or 10 bedrooms, bringing the price per room to something more reasonable, about $1250. (Except for the first listing here — if accurate, hopefully nobody signed a lease there!)

First 5 Listings with Posted Price > 12000

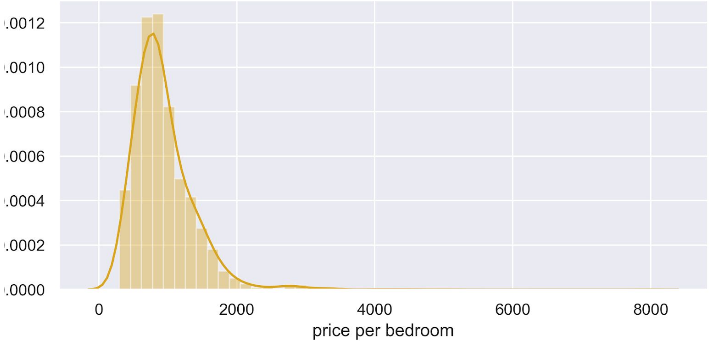


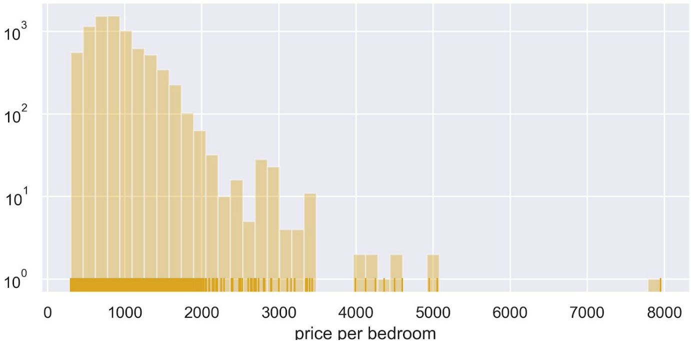
	address	price	bed	bath	area	company	neighborhood	laundry	pets	parking	utilities	property_type	year_built	description
1803	39 Menlo Park	15900.0	2.0	2.0	1056.0	None	None	1.0	0.0	1.0	[]	None	NaN	39 MENLO PARK: 1992 Redman. 2-bedroom 2-bathro...
3730	10 br, 3 bath House - 725 Church Street	12530.0	10.0	3.0	NaN	None	None	0.0	0.0	1.0	[]	apartment	NaN	(734) 663-8989 - 14 Bedroom House - Available...
5605	725 Church Street, Ann Arbor, MI 48104	12530.0	10.0	3.0	NaN	None	None	NaN	NaN	NaN	[]	house	NaN	
7822	723 Oakland Avenue Ann Arbor, MI 48104	12500.0	9.0	4.0	NaN	Bartonbrook Properties, LLC	East Packard	1.0	0.0	1.0	[Recycling, Trash Removal included]	house	NaN	NEW RENOVATIONS!! 9 bedroom house located in a...
7823	814 Hill Street Ann Arbor, MI 48104	14250.0	10.0	3.0	NaN	Bartonbrook Properties, LLC	East Packard	1.0	0.0	1.0	[Recycling, Trash Removal included]	house	NaN	Beautiful 12 bedroom/3 bath house, WITH UPDATE...

With this in mind, we plotted the price per bedroom, as this might be a more informative statistic.[1] (If you need two bedrooms for yourself, we won't judge.)

[1] Again, per-room prices below $300 might be an artifact of the data: it is definitely possible that some of the scraped listings already calculated the per-bedroom price. Therefore, we did not transform the price for those listings whose per-bedroom price would end up under $300.

Most per-bedroom costs, then, lay in the $500-$1500 range — reasonable. Regardless, there are a surprising number of listings — 670, or ~8.6% — which are over $1500, some even nearing West Coast prices (Seattle's average rent is $2139, according to rentcafe.com). Overall, housing *can* get quite expensive, but there are plenty of "affordable" options.



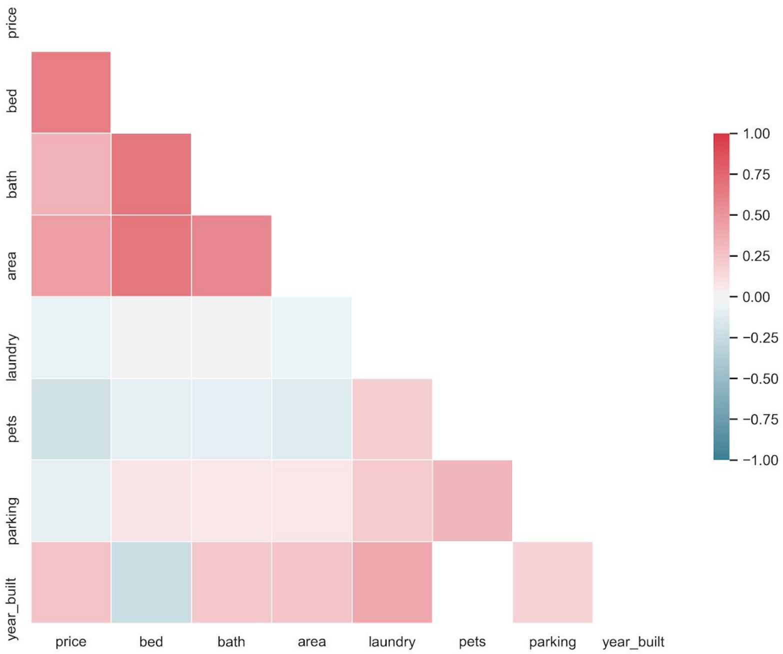


What trends can we observe when we sort by different features?

What features are most correlated with price?

We created a correlation heatmap to visualize pairwise correlations between our numerical and binary features.

Features such as number of bedrooms, number of bathrooms, and price seemed to be strongly correlated, which matches with intuition. Interestingly, there seemed to be a slight negative correlation between number of bedrooms and year built, and pets and price; the reason for this is unknown. However, the associated coefficients are relatively small in magnitude. Many of the other correlations are very weak.

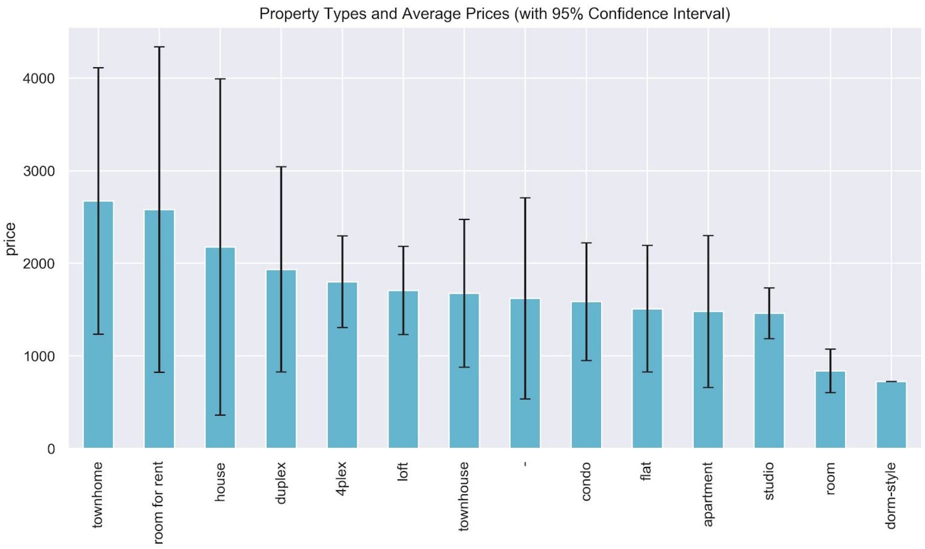


Does the price depend on what kind of unit you lease?

We tried to break down price by property type [2] to see if there was any notable difference in price based on what kind of unit was leased.

[2] Many of the categories are biased due to low sample size — we dropped all categories with only 1 representative entry, since it would be impossible to calculate the standard error of the mean. It should be noted that most off-campus students live in either houses or apartments.

Some of the point estimates differ, but there is so much variability that we cannot definitively conclude (with 95% confidence) that there *is* a difference, especially between houses and apartments. In fact, very few of the categories have non-overlapping bars. It doesn't look like property type affects price too heavily.

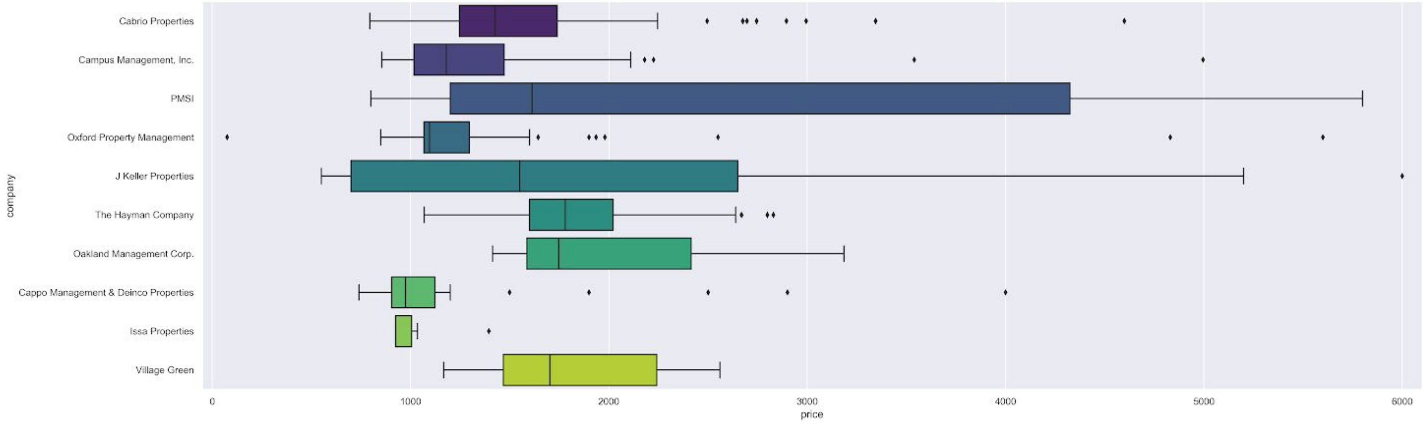


Are certain rental companies more expensive than others?

Look up any rental company, and you'll see that the vast majority of them have dismal ratings ranging from 1-2 stars (perhaps fairly, but also perhaps unfairly!). This is likely due to a selection effect, since tenants with pleasant experiences are generally less likely to post a review as opposed to someone more disgruntled. In any case, if you have a bad experience with a rental company, it might as well be cheap; so, we set out to analyze the distribution of prices by rental company, something that could still be explored in much more depth.

Distribution of prices: 10 most common rental companies.

Outliers with a price greater than $6000 were excluded from the graph for greater visibility. Issa Properties has a strikingly tight distribution, while PMSI has the largest spread.

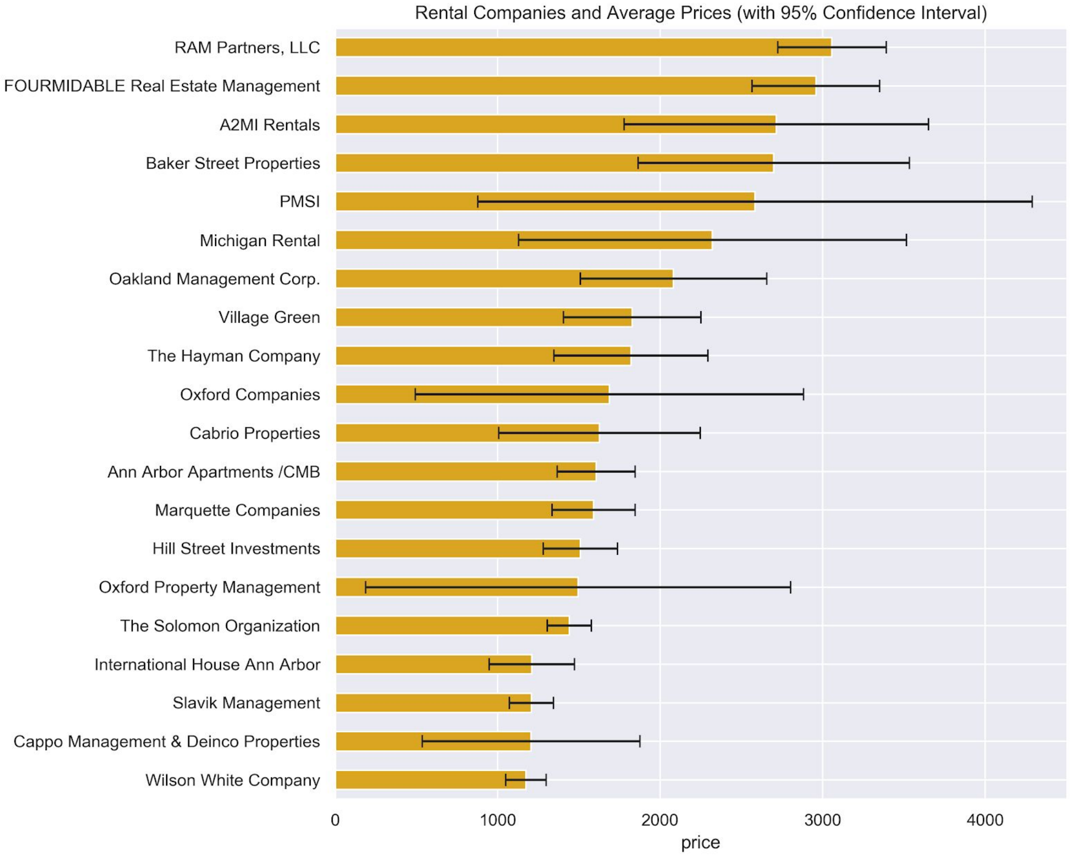


This was interesting, so we did more research. According to Issa Properties' website, it appears that they are a family-owned business that has "purposely remained small", with 3 available listings. On the other hand, PMSI's website states that they offer units "all around the Greater Ann Arbor area"; they seem to be a much larger company in general, being the 7th most frequent company in our dataset.

Average price: 20 most common rental companies, excluding missing prices.

Here, we wanted to see how average price (and variability) varied across different companies. This time, we took the 20 most-common [3] rental companies in our dataset (note that the below figure is NOT sorted by frequency; it is sorted in order of decreasing average price).

[3] Since we were computing statistics based on price, we chose to exclude all listings without a recorded price. This caused the most frequent companies to change from before. Noticeably, Issa Properties, our 9th most frequent company from the boxplot above, only had 14 entries out of 112 with recorded prices, causing it to disappear from the top 20 entirely.



There seems to be a lot less variability here than when sorting by property type; noticeably, there actually seems to be a significant difference in average price between properties owned by RAM Partners, LLC and those owned by the Wilson White Company, for instance. Students who are looking for cheaper housing may want to take this into consideration.

Location, location, location: how much does it matter?

Some of us lived on North Campus our freshman year and have seen firsthand the amount of hate it gets. Many of the hapless freshmen who get placed there move to Central Campus next year, never to return ("I might be guilty of that" — Eric).[4] Let's take a more objective look at this and break the data down at the neighborhood level.

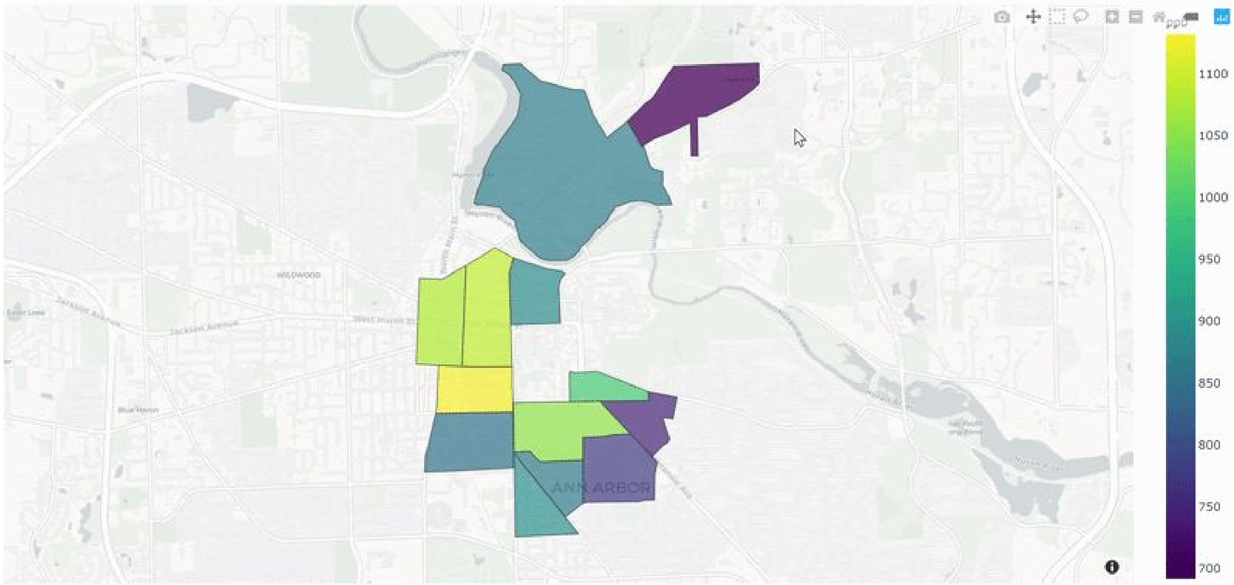
[4] It might be cool to collect data on *retention* — what proportion of people choose to stay on North Campus for more than one year?

The neighborhood map can be found in the repository, but is also posted on Michigan's off-campus housing site. Note that we elected to add another neighborhood named "Northside" in that weird region between North Campus and Central Campus, by Argo Livery.

Naturally, this section contains lots of maps, which are best viewed interactively; they can be found in the repository under `map/plotly-maps/` and viewed with this tool.

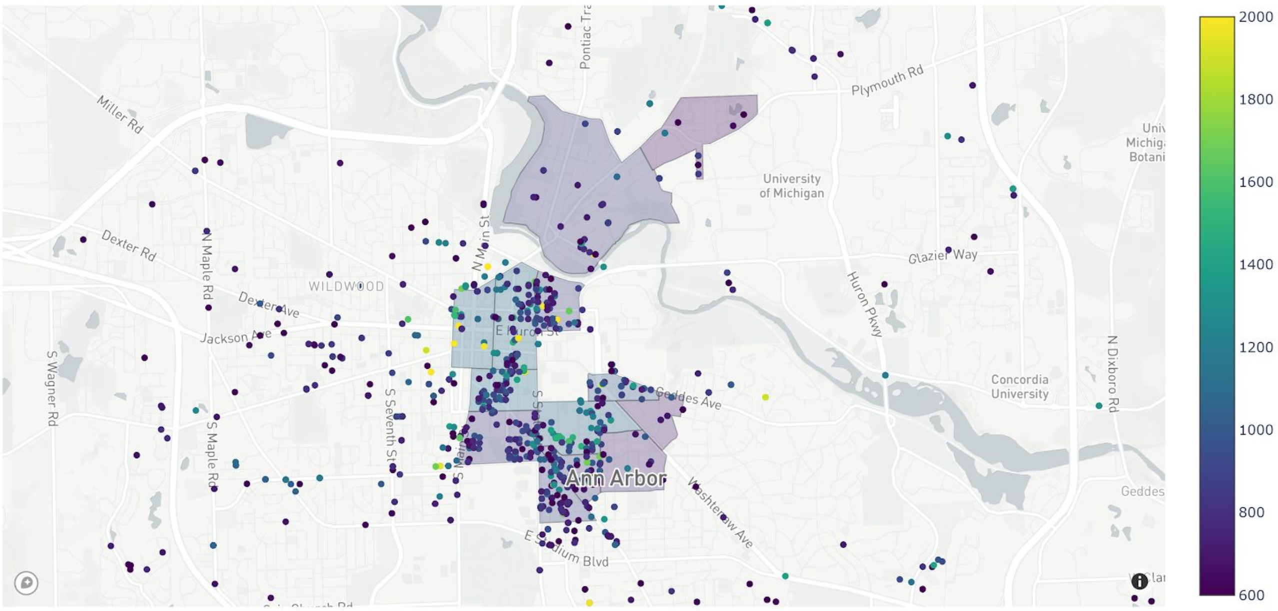
We built an interactive map using D3 and Plotly to combine most of our findings into one visualization, which can be viewed here. We summarize some of these findings below.

Average price by neighborhood. (link)



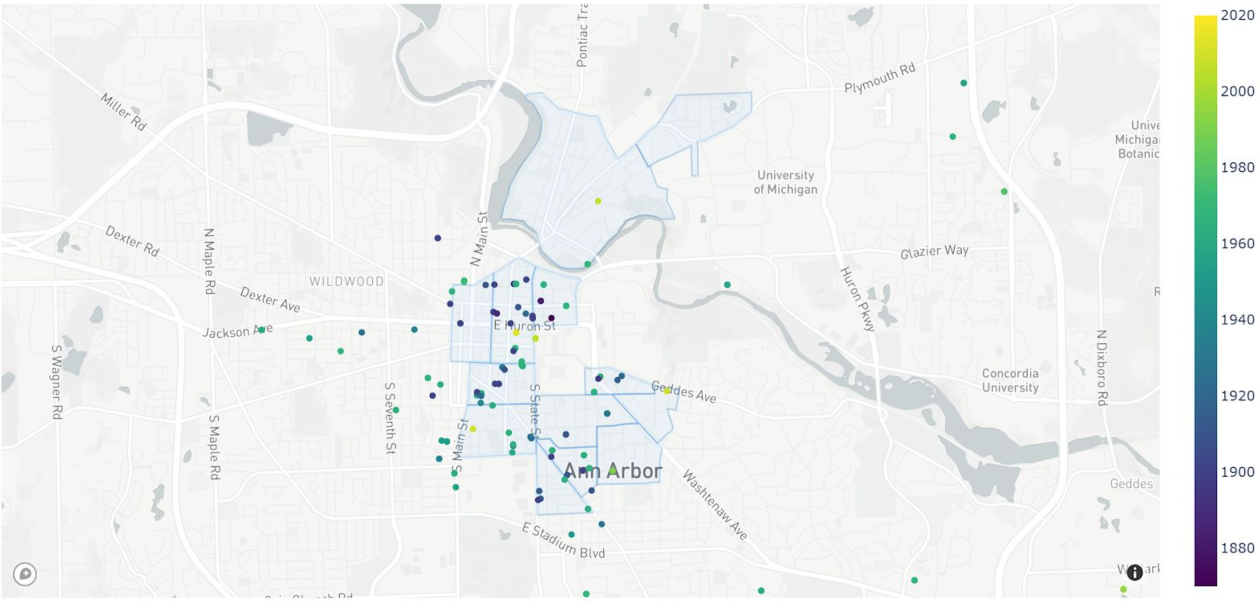
If there is one advantage to living on North Campus, it's the absurdly cheap prices (also, Kroger) when compared to properties down on Central Campus; this is likely to do with the higher demand down there. The most expensive neighborhoods seem to be in the downtown area, averaging over $1000 per room. With the exception of this area, it seems that neighborhoods further away from campus proper, such as Oxbridge and North Burns Park, are cheaper. This is something we'd expect, but it's nice to confirm it with data.

Distribution of listings, colored by price per bedroom. (link)



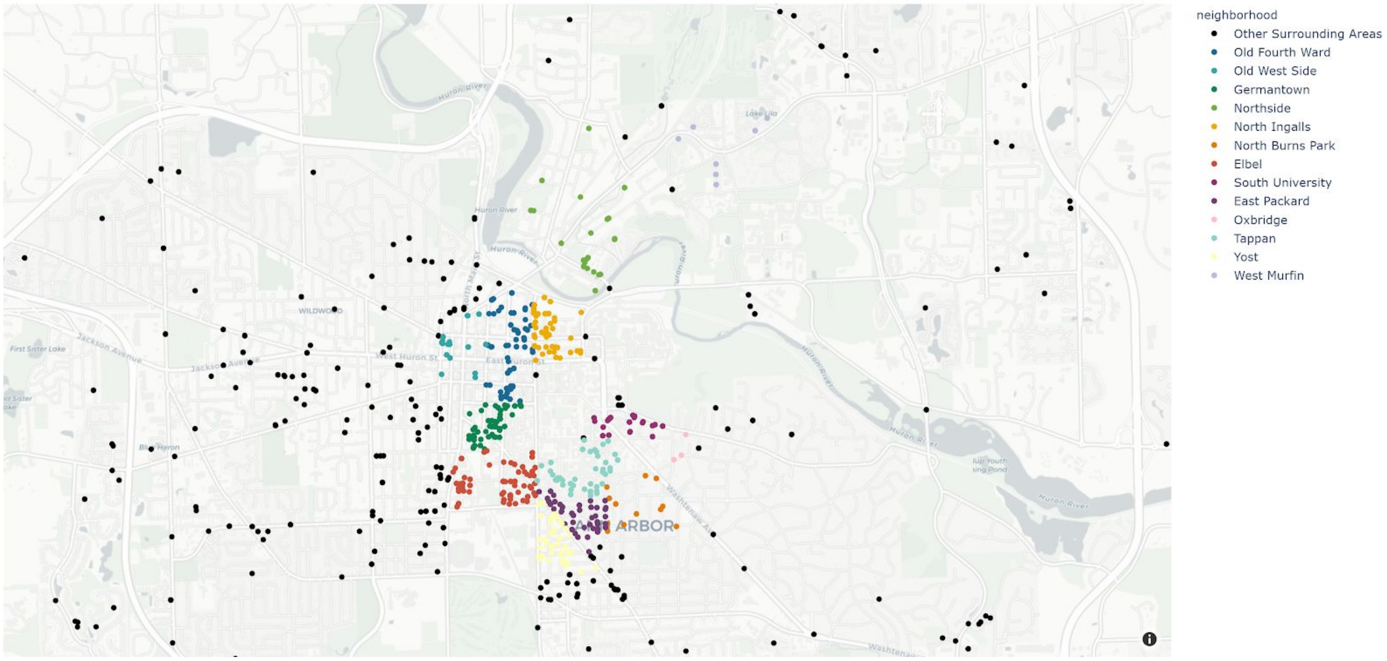
It seems that many of the most expensive properties lie in the heart of Ann Arbor; you can see a fair amount of listings with price ~$2000 in the Old Fourth Ward, North Ingalls, and the Old West Side. However, there is also a relatively dense cluster of ~$1500 listings in Tappan.

Distribution of listings, colored by year built. (link)

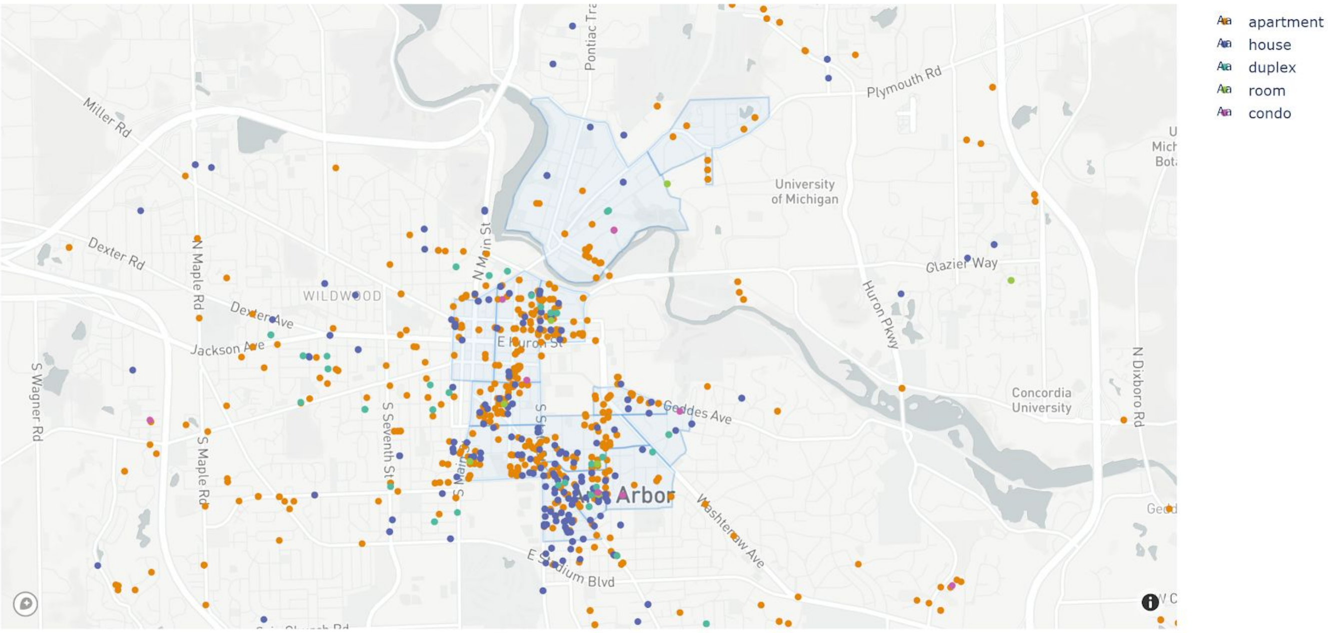


These data are relatively sparse, given that many of the sources we used did not provide us with such information. One cool thing to note — there are some properties that have been around since the late 1800s and are still available to rent!

Distribution of listings, colored by neighborhood. (link)



Distribution of listings, colored by property type. (link)



As an aside, I find this to be a very aesthetically-pleasing map. Just look at the colors!

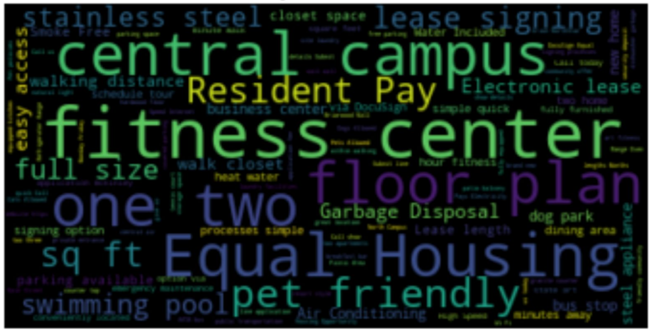
How are realty companies marketing their properties?

Most listings came with descriptions of the unit; what interesting information could we extract from that? We created word clouds to visualize the most frequent words and phrases in descriptions of listings, grouped by various criteria.

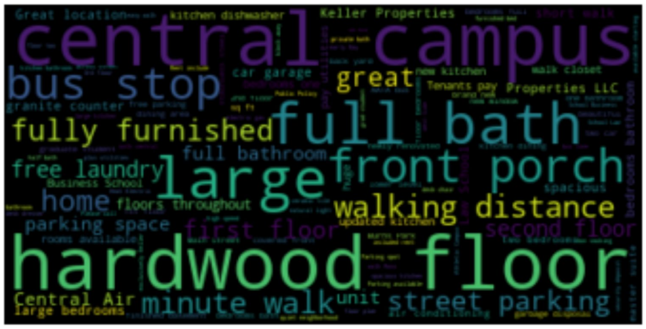
Descriptions by property type.

The vast majority of students live in apartments or houses on Central Campus; therefore, it makes sense that the phrase "central campus" appears so often in these listings. Noticeably, condos emphasize access to the Ann Arbor city bus ("the Ride"), as well as having "steel appliances"; these listings are likely targeted at non-students in the housing market.

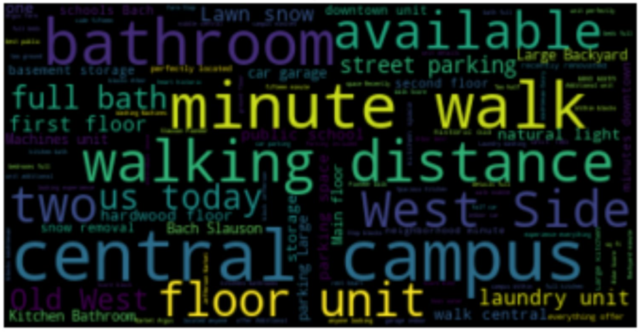
Apartment



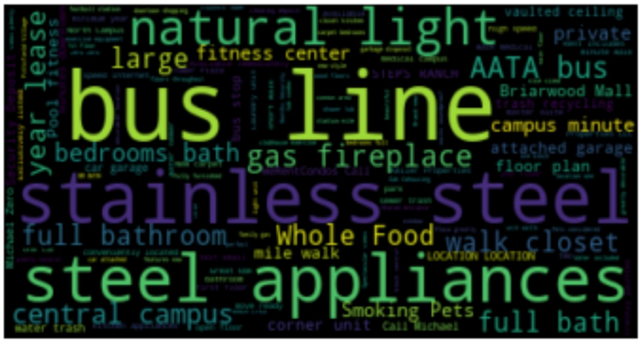
House



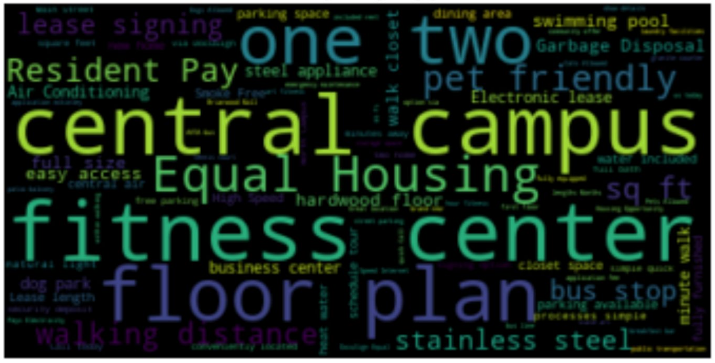
Duplex



Condo



All Property Types

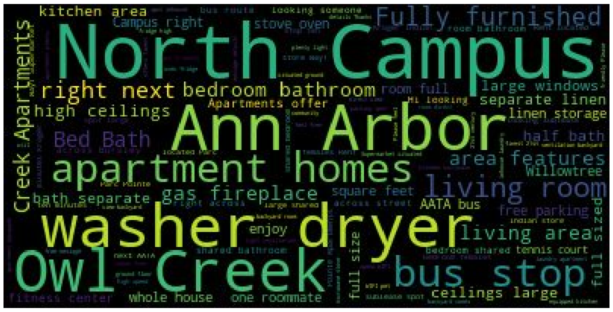


Descriptions by neighborhood.

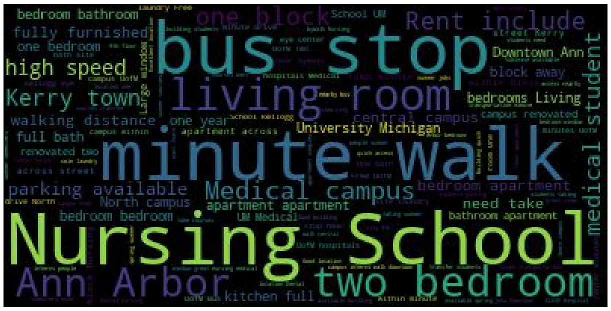
There are distinct differences in the descriptions at the neighborhood level, most based on nearby points of interest. This report only includes some of the figures; the rest of them are located in `NLP/Clouds/Neighborhoods/`.

Listings in the North Ingalls neighborhood seem to target nursing students, and listings in West Murfin of course emphasize its North Campus location, as well as proximity to a bus stop.

West Murfin

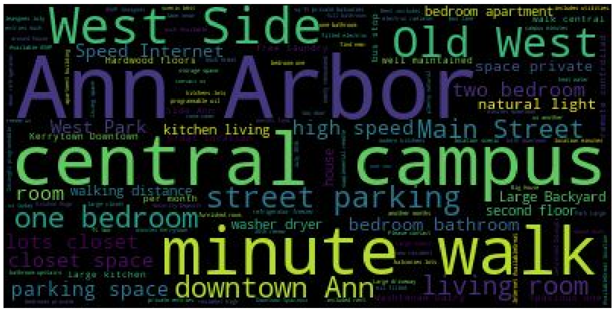


North Ingalls

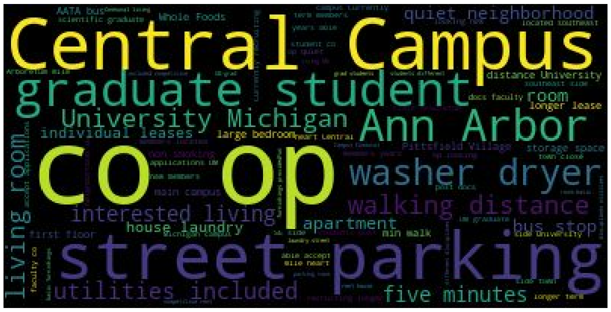


Interestingly, "co-op" seems to be one of the most common words in the descriptions of listings located in the Oxbridge neighborhood; there are 5 co-op houses in the area, 2 in Oxbridge itself and 3 on Washtenaw Avenue's other side, in North Burns Park. Listings on the West Side of Ann Arbor, while farther from campus, seem to advertise downtown a little more than the rest.

Old West Side

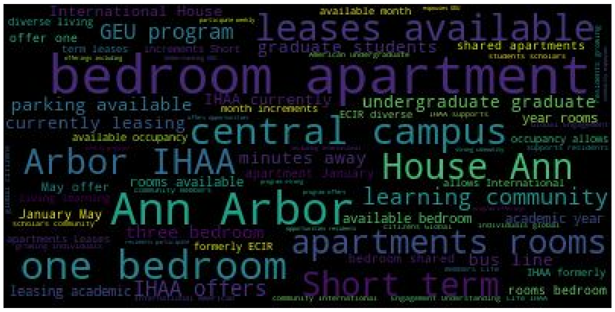


Oxbridge

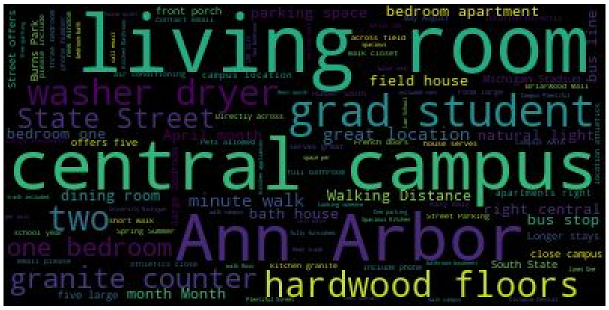


Yost is one of the only neighborhoods where the phrase "grad student" is frequent enough to even show up in the cloud; these listings also seem to be quite proud of their hardwood floors. Tappan's listings are somewhat interesting; International House Ann Arbor (IHAA) and its learning community, Global Engagement for Understanding (GEU), are mentioned by the eponymous organization frequently enough to show up in our word cloud of 440 listings.

Tappan

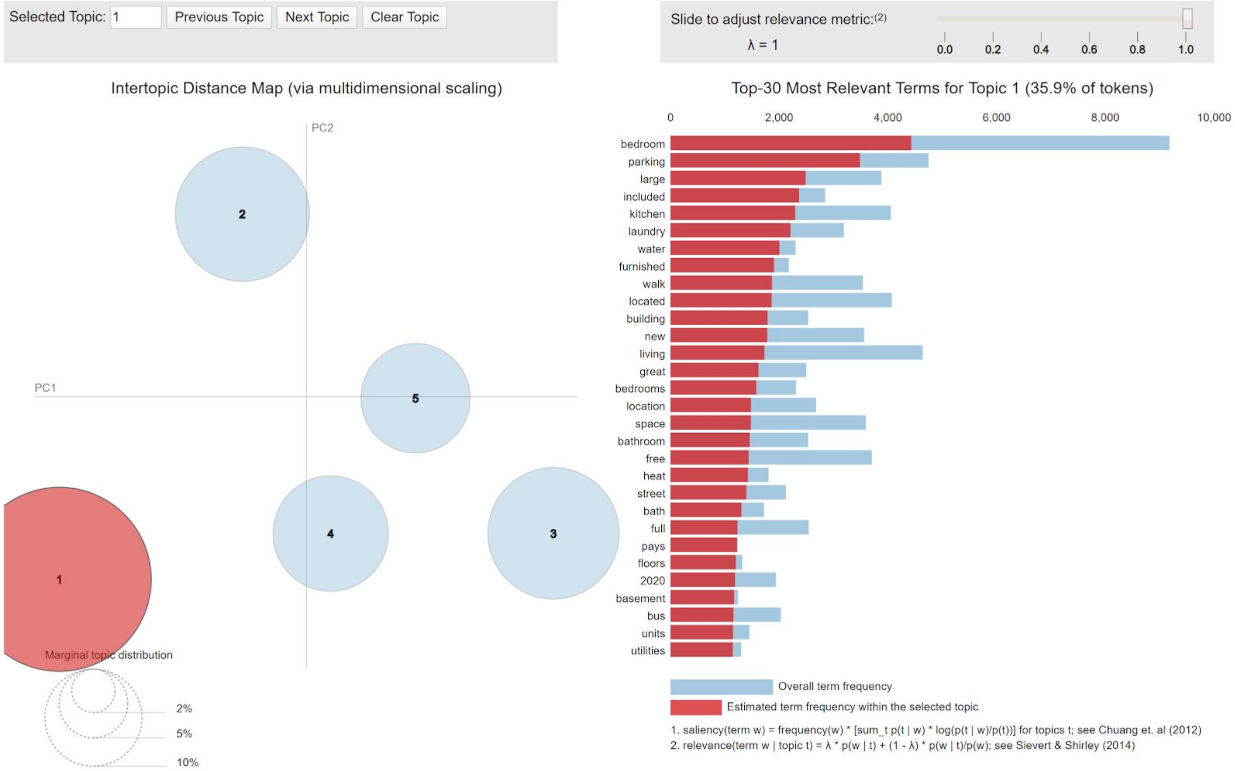


Yost



Topic Modeling.

We used a generative statistical model called latent Dirichlet allocation to automatically find "topics" in the descriptions; these topics were visualized using pyLDAvis, a Python wrapper for the original R library. The results varied greatly depending on the number of topics k; you can find the interactive visualizations in `NLP/LDA/`. The 5-topic model can be viewed here.


Selected Topic: 1

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Each topic consists of a group of words, ranked by relevance. The topics are a little hard to interpret, but some noticeable findings include:

- Topic 1 seems to contain all the common words you'd expect to see in a description: "bedroom", "parking", "kitchen", etc; the other topics seem to contain much more specific terms. This is something I'd expect to see out of a text generator like this.
- The most relevant terms for Topic 2 seem to be about housing tours and the leasing process, with words such as "application", "call", "info", "show", and "signing". It also appeared to pick up on contact information, with terms like "online", "electronic", and "wwwmckinleycom" — obviously a website — showing up.
- Topic 3 appeared to emphasize amenities or perks, with high-ranking terms such as "center" — presumably referring to fitness centers, although there are other possibilities — "private", "amenities", "fitness" itself, "pool", "access", and more.
- The term "bedroom" was the top term in Topics 1, 2, and 4; it still appeared in the top 30 for Topics 3 and 5. This is especially interesting later on, when we find out that number of bedrooms is the most influential feature in predicting price!

Note that these findings were for the 5-topic model; given that we could not find a discernible trend for Topics 4 and 5, it is unlikely that increasing k would garner much more insight.

The danger in this is the interpretation — since LDA is an unsupervised model, it is likely that we are succumbing to a bit of confirmation bias, in that we are actively searching for patterns in the resulting topics, when there might not exist any! Thus we should exercise caution.

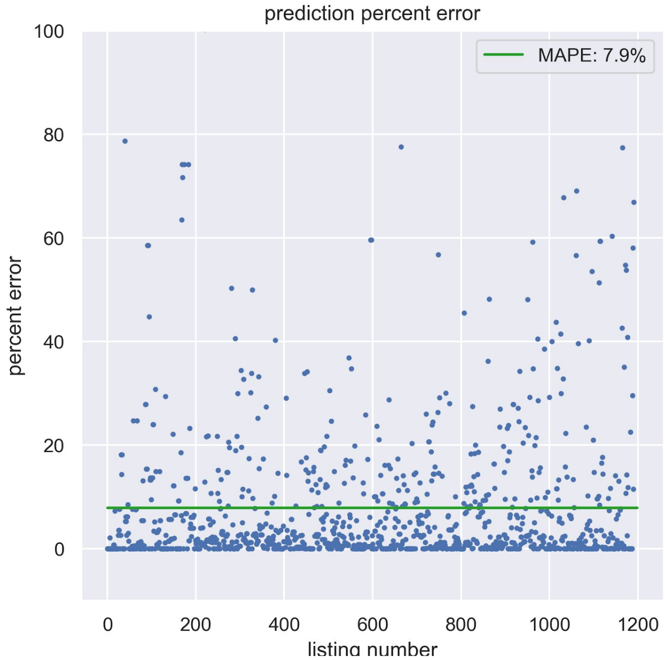
What features can be used to predict price?

Students looking for off-campus housing often go in with an idea of what they are looking for — who they want to live with, a general location, and other such "requirements", as we'll call them. We sought to develop a regression model to predict the price of a hypothetical listing with these requirements, in hopes that it could become a usable tool for students who have such preferences and would like a reasonable price estimate.

We restricted our data to houses and apartments; we assumed that most students would be looking for housing of this type. We also dropped most categorical features such as description, images, and address; one-hot encoding all these features would be counterproductive due to increased dimensionality (see discussion below), and they would not be typical "requirements" anyway. Finally, we dropped very sparse fields such as `year_built` (~81.1% sparsity) since it's unlikely that their inclusion would have increased the model's predictive power.

Regression model and performance.

After data cleaning, we trained a random forest regressor on about 3800 listings' worth of data and achieved a test mean absolute percent error (MAPE)[5] ranging between 7-9% on the exponentiated predictions. This means, in the absence of some sort of bias,[6] the model's prediction would be on average ~7-9% away from the true price.



[5] We chose MAPE as our test metric as opposed to something more traditional, like mean squared error (MSE), since we figured error was relative and should be perceived as such. For instance, an error of $200 is not that large compared to a property that costs $1000, but is a lot more significant when the property costs $500. To deal with this, we log-transformed the price before training the regression model so that a difference in log(price) actually corresponded to a relative error between the true price and predicted value; this way, we could still train the model using MSE as our criterion.

[6] To check for such a systematic bias (e.g., our model performs well on cheaper listings but does worse on more expensive listings), we sorted the percent error in order of increasing true price before plotting it. That specific bias, at least, does not seem to be present, although the model does appear to do worse at either end of the price range than in the middle.

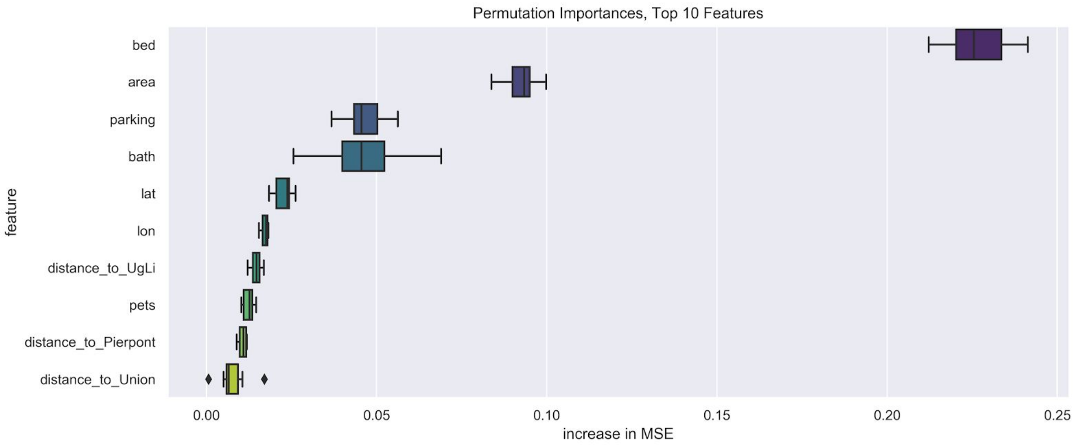
Feature importance and interpretation.

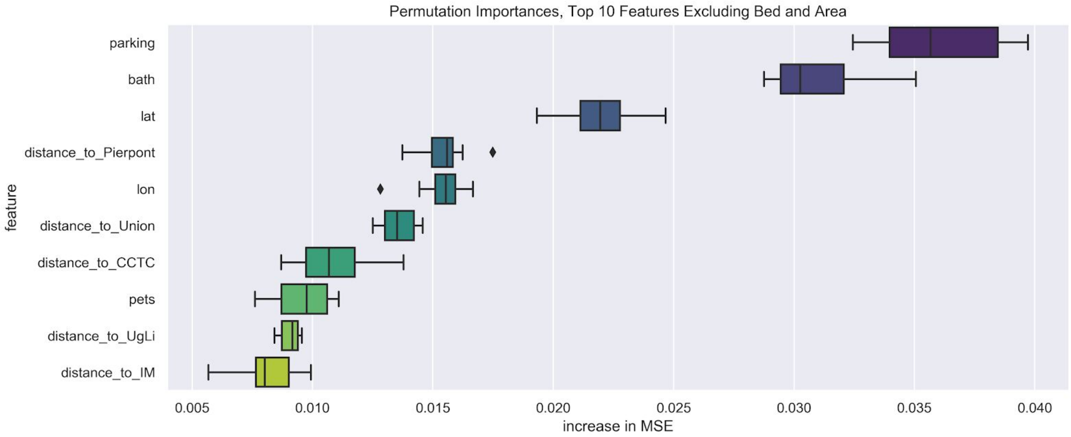
Random forests are hard to interpret on their own, so we sought a way to extract feature importances from our model and see which features would give us the most insight. We implemented a couple of methods, including one that used concepts from game theory!

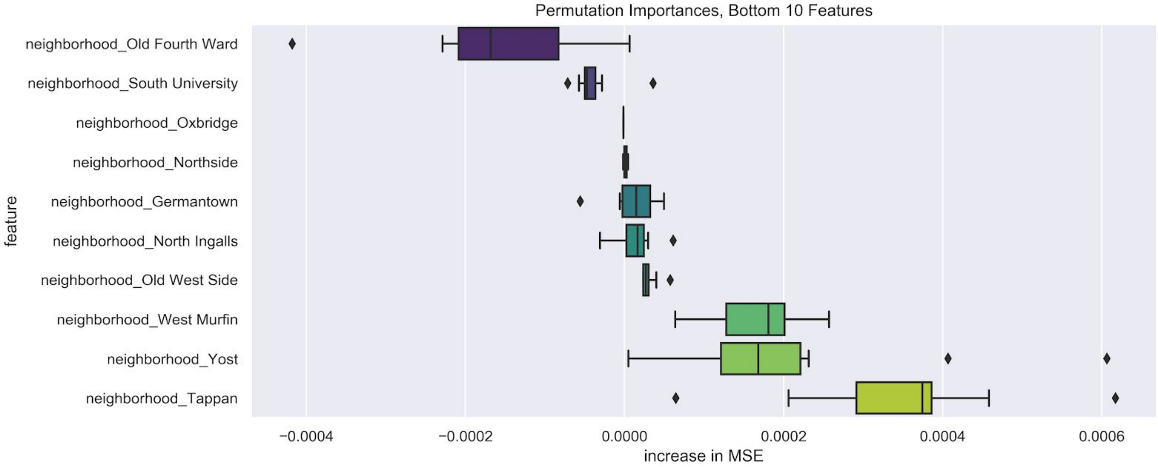
Permutation importance.

We used a method called permutation importance [7] (Breiman, 2001) to measure the relative predictive power of each feature. We found that the number of bedrooms was by far the most informational feature; the additional loss incurred when randomizing this feature was more than double that of the second-most significant feature. The test MSE itself was ~0.031; therefore, it looks like randomizing some of these features had a drastic effect on performance! In particular, the loss increased by almost 700% when the number of bedrooms was permuted.

[7] It has been argued that permutation importance is a better way to measure feature importances than the traditional Gini impurity-based metric, which is biased towards numerical variables; see here.







All of the neighborhood features were ranked at the bottom; some of them actually had a *negative* effect on the model's performance, meaning that their inclusion was counterproductive! However, this may be misleading due to the nature of one-hot encoding.

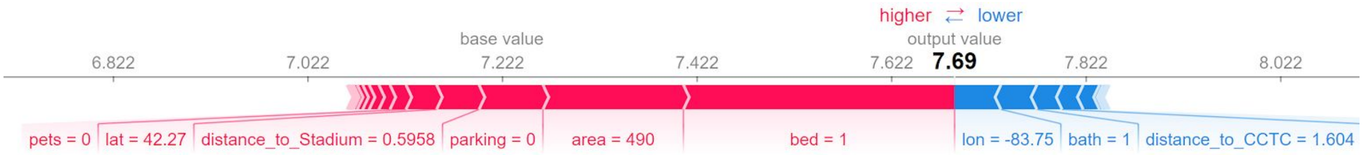
Following from above, a surprising thing to note is that scikit-learn's decision tree models cannot work with raw categorical data; they must be encoded in some way. We chose to one-hot encode our categorical features, but this makes it harder to extract the importance of `neighborhood` or `property_type` as standalone features, in addition to inducing artificial sparsity in our model — 16 out of 33 features were indicator variables. Thus we cannot precisely state the influence of each categorical field as a whole. In the future, it may be worthwhile to try using a different implementation of the random forest regressor (R's `randomForest` class comes to mind) to make predictions and compute permutation importance.

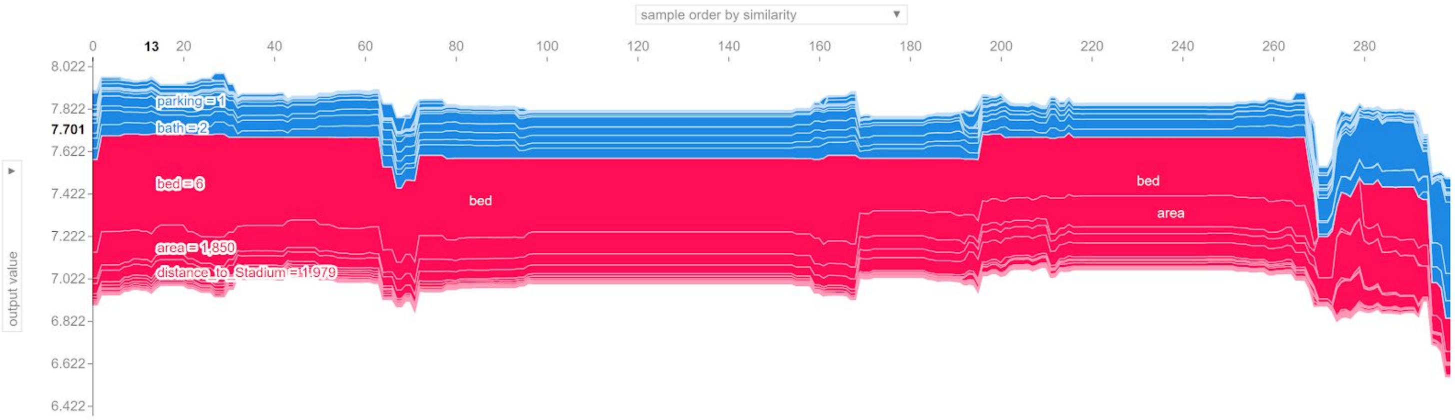
SHAP: a game-theoretic approach.

We sought to try another method to see how it compared to the permutation importance from above. A 2017 paper by Lundberg and Lee, published in NeurIPS, extends a concept from game theory known as the Shapley value to measure feature importance.

Intuitively, if we treat each feature as an agent in a cooperative game and define the difference between the prediction and the average prediction over all data points to be the payoff or "gain", then the SHAP (Shapley Additive Explanation) value is a unique distribution over the features consisting of each one's contribution to the gain — this is exactly variable importance, and should seem quite similar to the permutation importance method! The Shapley value in game theory is an "optimal" distribution of sorts and possesses many desirable properties, but those are beyond the scope of this report.

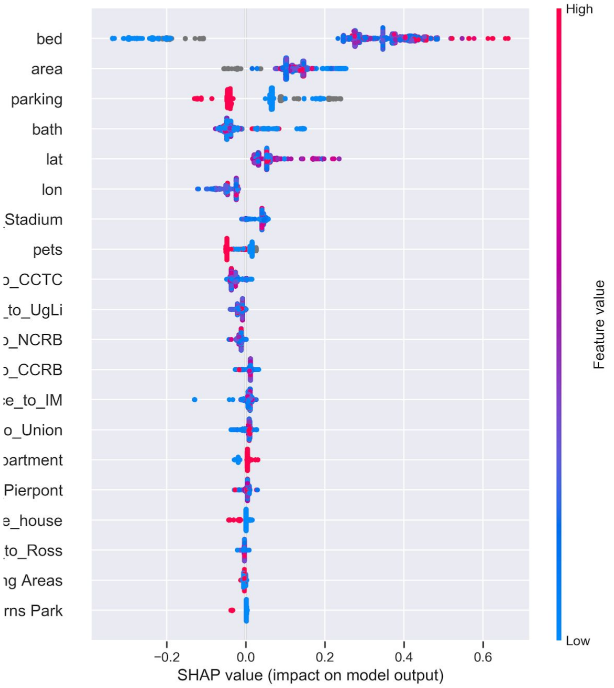
In the visualizations below, features were colored in red if they had a positive effect on the prediction and colored in blue if they had a negative effect. The *force plot* on top shows the distribution for one such listing; the size of each arrow corresponds to the magnitude of that feature's effect. The interactive streamgraph on the bottom shows the distribution over a random sample of 300 listings, which can be explored in more detail here.

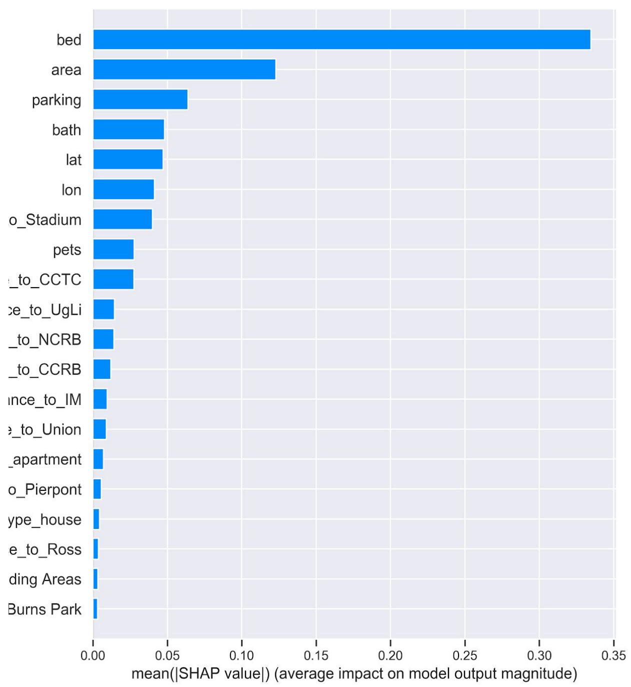




We can see that the number of bedrooms is by far the most important feature, just as we observed before. Noticeably, there are a couple of listings where the number of bedrooms is not as important; these data points seem to be missing this feature.

Grouping the SHAP values by feature, we can clearly see that our results here correspond closely to those found through permutation importance — the top 10 features are ranked in almost the exact same order, with a Kendall rank correlation coefficient of ~0.911. The figure on the left also shows the direction of each effect; noticeably, the presence of parking seems to *decrease* price, which is counterintuitive.

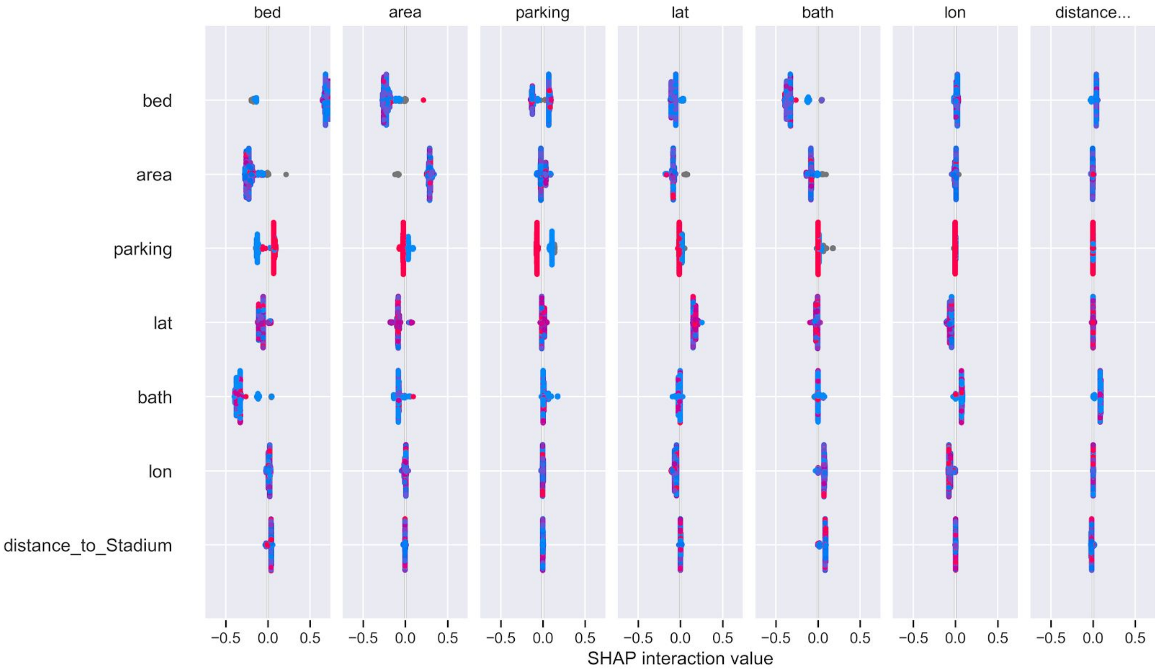




It may be worthwhile to rerun the model with actual prices instead of log-transformed prices as labels; this way, we could more tangibly see the effect each variable has on the price itself.

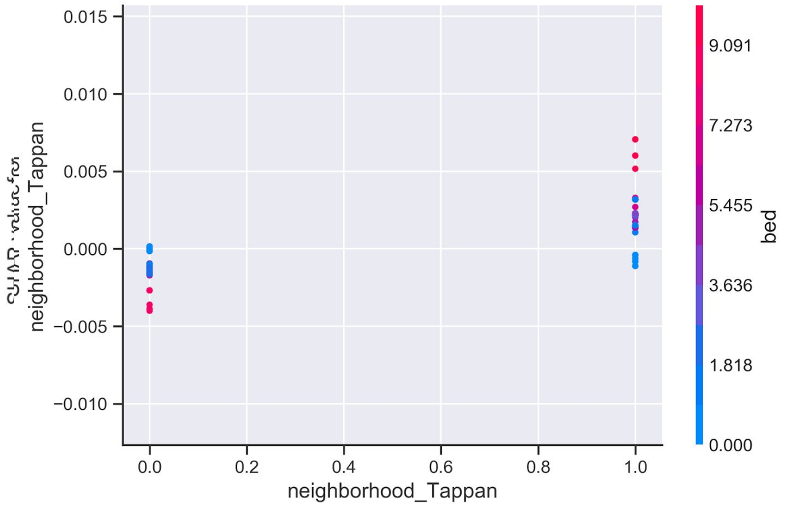
Feature interactions.

We also wanted to investigate any potential feature interactions. For instance, it's possible that parking availability could change based on neighborhood or property type. The interactions are given by the off-diagonal entries. From the chart, it doesn't look like many of the feature pairs have a discernible interaction term, but some interesting ones are highlighted below.

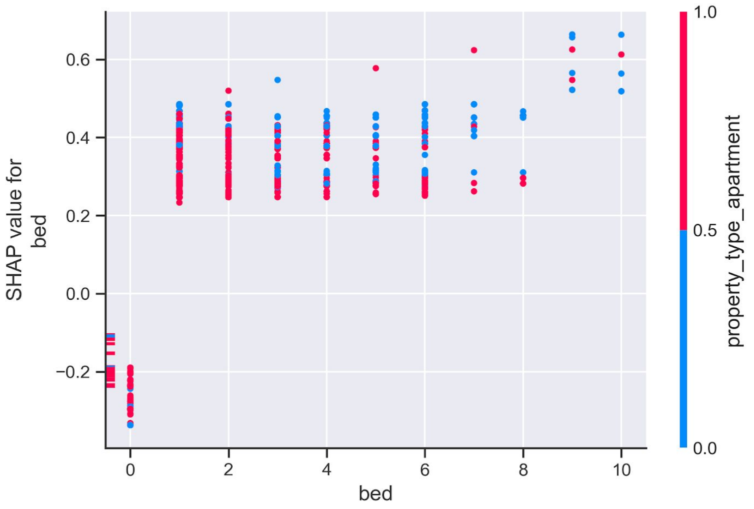


This plot shows the importance of the Tappan indicator feature, colored by the number of bedrooms. It looks like having more bedrooms actually pushes the price down if the property is outside of Tappan, but is reversed when the listing is located in the neighborhood. This does not make intuitive sense, since the overall trend was that more bedrooms implies a higher price.

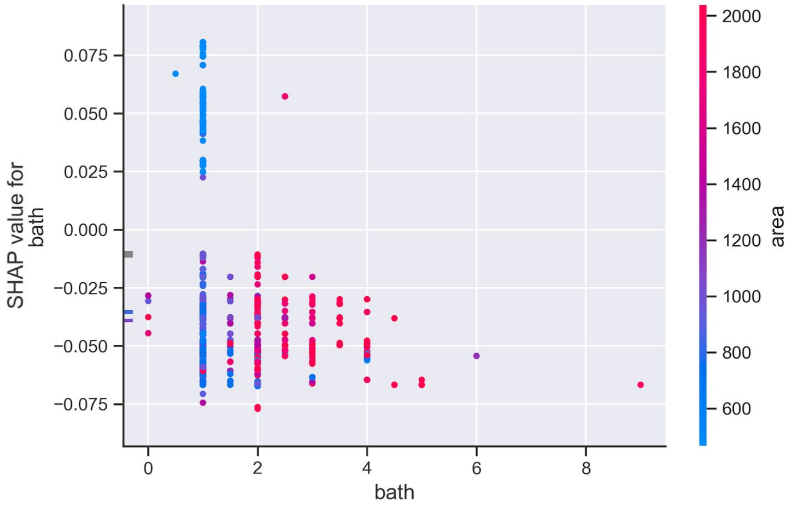
This graph may be misleading, as some points look to be obscured; alternatively, this could be a side effect of small sample size.



Here we can see how the SHAP value of the number of bedrooms varies by property type. There doesn't appear to be a distinct trend in SHAP value, but the graph shows that apartments tend to have less bedrooms. This may be a reason why apartments seem to be a little cheaper than houses overall.



Interestingly, if we exclude all 1-bathroom listings, there appears to be no real association between the number of bathrooms and area; one would expect a positive relationship. Having more than one bathroom actually pushes the price *down* for almost all of the relevant data points; it is unknown why this is the case.



From these plots, it is hard to tell if there are any true, strong interactions; these look spurious at best, but it may be worth digging into the data and investigating further. Note that we may get better results with a model that can handle categorical data without one-hot encoding.

4. CONCLUSION.

So, what have we learned?

Despite popular opinion, there *does* exist affordable housing in Ann Arbor — at first glance. However, when you start to filter by factors like bedrooms, bathrooms, and location, the search space can be vastly narrowed.

It's likely that factors outside of those encapsulated by our dataset contribute to this perception of expensive housing; for one, there is definitely a time-sensitive component to the housing search as people rush to sign leases in the fall. The recent construction of luxury high-rises may influence people's perceptions as well.

If you're a student looking for cheaper housing, there are some tangible things you can do. For instance, some rental companies seem to have cheaper properties on average, and living farther from campus might also lower the price. However, it is likely that you'd have to make some concessions. From personal experience, we recommend simply starting the housing search early so properties don't fill up!

From a more research-oriented standpoint, there were quite a few interesting findings that surprised us. We found the description section rather insightful; there was a discernible difference in what features rental companies chose to emphasize from neighborhood to neighborhood, as well as property type, which revealed some cool insights into the target market. It would be worth investigating this further.

Another one of our goals was to see how we could use our data to predict price. We were surprised by how well our regression model performed on such noisy data, although there were some major outliers. It's not surprising that the number of bedrooms was the single largest predictor of price, but some of the other predictors, including neighborhood, seemed to have counterintuitive effects. For instance, despite there being a difference in average prices by neighborhood, our random forest model saw these features as the least important. We also expected more feature interactions, but that may have been hampered by the encoding strategy or small sample size.

Overall, there are still a lot of questions to be explored and data to be collected, but we hope this project has shed some light on the state of off-campus student housing in A2 today.

Addendum: One of the main motivators behind this project was to help inform Michigan students about the housing search. In the fall, we plan to publish a (summarized and decidedly less technical) article in the Michigan Daily's *Statement*, which will be complemented with a (unsummarized and highly technical) blog post on MDST's website. This report will be updated with a link to said articles when the time comes; until then, these links will be in red.

5. LIMITATIONS AND FURTHER RESEARCH.

From the beginning, we were surely limited by the nature of our dataset. Many websites included fields that were not found on other websites, making the data too sparse to be of use; for example, construction date (`year_built`) was only listed on ShowMeTheRent. Problems with missing data and incorrect values also caused about half of the original listings to be unfit for training/testing the regression model. However, these data are still valuable; we encourage readers to check out the repository and conduct their own analyses with the existing data.

Of course, there are many other possible factors that were not available from the sources we used. This includes tenant reviews, historical listing data, and even the current distribution of students living off-campus. There are also many more open properties in the fall, and collecting data during the housing rush may lead to different results. The following are some potential areas of further exploration:

- If data can be obtained on historical rent prices and listings, it may be worthwhile to analyze the *change* in housing prices over the years.
- How does Ann Arbor measure up in comparison to other college towns? Is it more expensive on average than other cities sharing similar characteristics?
- How does on-campus housing (i.e., dorms) compare to off-campus housing when considering factors such as food cost, utilities, and parking?

On the administrative side of things, we were also constrained by time, messy data, and of course, the COVID-19 pandemic. Data cleaning took up the vast majority of our work time, and remote work sessions were not exactly easy to set up! Special thanks to those who decided to stick around after poop hit the fan.

6. EXTRAS.

Libraries used.

General:	NumPy, pandas, matplotlib, seaborn
Web scraping:	BeautifulSoup, regex
Geocoding:	GeoPy, GeoPandas, Shapely
Mapping:	Plotly, D3
NLP:	wordcloud, pyLDAvis
Regression:	scikit-learn, SHAP

Stretch goals.

1. Collect historical data and show trends over time
2. Bundle maps into full, interactive web app
3. Classification/regression using image data
4. Comparison to other college towns