Data Immersion Task 6.1 - Sourcing Open Data

Analysis of affordability and property availability (both rental and for sale) in US areas with major food production facilities.

This project uses multiple different publicly available data sources:

1. Historical Monthly Inventory Zip data from Realtor.com
2. Zillow Observed Rent Index by ZIP code
3. Apartment List Vacancy Index
4. Apartment List Rent Estimates
5. Zip code list acquired from major protein company facility locations
6. Realtor.com scraped homes for sale listings for zip codes in list
7. Realtor.com scraped rental listings for zip codes in list

1. Historical Monthly Inventory data from Realtor.com

Data Source:

This open source data set is publicly available. It was downloaded from realtor.com on their data page: https://www.realtor.com/research/data/ The "Zip" "Historical Data" was chosen. Future usage will add "Current Month Data" every month.

Data Collection:

According to realtor.com, this data "is based on the most comprehensive and accurate database of MLS-listed for-sale homes in the industry. We aggregate and analyze data from hundreds of sources and produce hundreds of metrics for multiple markets, and curate figures and trends where possible for reliability and comparability."

Data Contents:

This data includes historical real estate data aggregated by zip code on a monthly basis, including details such as: active listing count, average listing price, days on market, median list price per sqft, median listing price, median listing sqft, weekly new listing count, total listing count, month over month and year over year numbers for many of those metrics, and several other factors to calculate desirability or market demand.

Why this Data:

This data was chosen because it can give us an idea of how many homes are typically available in a given location, and what seasonal trends by location look like. It can also give us a baseline for considering average and median home price trends that can later be calculated to determine how many homes were likely available historically below a certain price point, that can then be used (together with current non-aggregated data) to forecast how many homes might be available below a certain price point in the future.

Data Profile:

The data originally had 40 columns and 2,298,453 rows. After consistency checks and cleaning it has 18 columns and 2,298,452 rows.

For cleaning I just dropped columns that won't be needed for this project, including all the month-over-month information. There was one row that was informational only and it was removed.

Columns	Description	Time Variant / Invariant	Data Type
Month_date_yyyymm	This has the month and year that the data was collected/compiled	Variant	Quantitative
Postal_code	The zip code where properties are listed	Invariant	Qualitative
City state	The city and state associated with the property listings	Invariant	Qualitatve
Median_listing_price	The median listing price within the specified geography during the specified month.	Invariant	Quantitative
Median_listing_price_yy	The percentage change in the median listing price from the same month in the previous year.	Invariant	Quantitative
Active_listing_count	The count of active listings within the specified geography during the specified month. The active listing count tracks the number of for sale properties on the market, excluding pending listings where a pending status is available. This is a snapshot measure of how many active listings can be expected on any given day of the specified month.	Invariant	Quantitative
Active_listing_count_yy	The percentage change in the active listing count from the same month in the previous year.	Invariant	Quantitative
New_listing_count	The count of new listings added to the market within the specified geography. The new listing count represents a typical week's worth of new listings in a given month. The new listing count can be multiplied by the number of weeks in a month to produce a monthly new listing count.	Invariant	Quantitative
New_listing_count_yy	The percentage change in the new listing count from the same month in the previous year.	Invariant	Quantitative
median_listing_price_per_square_foot	The median listing price per square foot within the specified geography during the specified month.	Invariant	Quantitative
median_listing_price_per_square_foot_yy	The percentage change in the median listing price per square foot from the same month in the previous year.	Invariant	Quantitative
median_square_feet	The median listing square feet within the specified geography during the specified month	Invariant	Quantitative
median_square_feet_yy	The percentage change in the median listing square feet from the same month in the previous year.	Invariant	Quantitative
average_listing_price	The average listing price within the	Invariant	Quantitative

	specified geography during the specified month.		
average_listing_price_yy	The percentage change in the average listing price from the same month in the previous year.	Invariant	Quantitative
total_listing_count	The total of both active listings and pending listings within the specified geography during the specified month. This is a snapshot measure of how many total listings can be expected on any given day of the specified month.	Invariant	Quantitative
total_listing_count_yy	The percentage change in the total listing count from the same month in the previous year.	Invariant	Quantitative
quality_flag	Triggered ("1") when data values are outside of their typical range. While rare, these figures should be reviewed before reporting.	Invariant	Qualitative

Limitations:

There are a lot of missing values in the new listing count year over year, and with a lot of the year over year data generally. So that data may not be useful when it's time to analyze it. However, the main limitation with this data lies in the fact that it's aggregated, but doesn't have any information about the portion of houses which were available within certain price ranges.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's presented publicly and all the information is aggregated without any possible personally identifying information.

2. Zillow Observed Rent Index by ZIP code

Data Source:

This open source data set is publicly available. It was downloaded from zillow.com on their Housing Data page: https://www.zillow.com/research/data/ The "Zillow Observed Rent Index" by "ZIP Codes" was chosen.

Data Collection:

According to Zillow, this data is collected by "calculating price differences for the same rental unit over time, then aggregating those differences across all properties repeatedly listed for rent on Zillow." This is intended to smooth out fluctuations in different costs of different unit types that are available at different times. According to Zillow, the data which these calculations are based is "Zillow's industry-leading database of rental properties. To ensure we are capturing the rental market as a whole, and not just those rentals posted to Zillow, we created weights for the index based on the latest data from the U.S. Census Bureau."

Data Contents:

This data includes historical ZORI (Zillow's calculated market rate rent) dollar figures for 6376 zip codes in the US on a monthly basis. It also includes the county, metro, city, and state names for those zip codes as well as how they're ranked by size within the list.

Why this Data:

This data was chosen because it can help identify trends in rental housing costs for a given zip code. Where those zip codes overlap with zip codes from our food production facility lists, it can be used to compare the ZORI rate to the observed average and median rates from currently available rental properties. This should help identify what future rental prices are likely to cost, which, when taken together with number of properties currently available for rent below a certain price point, might help predict future affordable rental property availability as well.

Data Profile:

The data originally had 106 columns and 6376 rows. After consistency checks and cleaning it has 102 columns and 6376 rows.

There were no duplicates, but I removed the columns for "RegionID", "SizeRank", "RegionType" because they won't be helpful to my analysis and "StateName" because it's basically a duplicate of the "State" column.

Columns	Description	Time Variant / Invariant	Data Type
RegionName	The zip code of each location	Invariant	Qualitative
State	The 2-letter state abbreviation of each location associated with the zip code	Invariant	Qualitative
City	The city of each location associated with the zip code	Invariant	Qualitative
Metro	The metropolitan area of each location associated with the zip code	Invariant	Qualitative
CountyName	The county of each location associated with the zip code	Invariant	Qualitative
2015-03-31	The ZORI (Zillow Observed Rent Index [rent rates]) for a given location during the month and year of the column name	Variant	Quantitative
One more for each month until	There are nearly 100 of those columns	Variant	Quantitative
2023-03-31	The ZORI (Zillow Observed Rent Index [rent rates]) for a given location during the month and year of the column name	Variant	Quantitative

Limitations:

There are a lot of missing values for various zip codes, especially the older you go (up to 80% of them have missing values back in March 2015). This gives significantly less data for trying to accurately extrapolate future rates. Secondly, there are a limited number of zip codes within the US for which this data exists at all, which means it's less likely to have all of the zip codes that I'm analyzing for this project.

Also, because this is a calculated number, there's always the possibility that the calculations aren't appropriate or have too many inaccuracies in some areas.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's presented publicly and all the information is aggregated without any possible personally identifying information.

3. Apartment List Vacancy Index

Data Source:

This open source data set is publicly available. It was downloaded from apartmentlist.com on their Data & Rent Estimates page: https://www.apartmentlist.com/research/category/data-rent-estimates The "Apartment List Vacancy Index (Jan 2017 - Present)" was chosen.

Data Collection:

According to Apartment List, the data "is calculated as the ratio of vacant units to total units among properties that list on our platform." Apartment List also describes multiple methods that they use to wait until a given platform's vacancy rates have generally stabilized before including it in their calculations. The data set only includes locations with at least 25 properties in each individual month over the course of the last year. However, their average monthly sample per city "consists of more than 17,000 units," indicating that only larger metropolitan areas are available in this data set.

Data Contents:

This data includes monthly historical rental property vacancy rates for 455 locations in the US. It also includes the names of the state and metro (where applicable), and the population of each location.

Why this Data:

This data was chosen with the hope that it might be able to help identify the historical number of available rental units within a given location (to be matched with the zip codes from the food production facilities list). At the very least, it might be able to help identify trends in increasing or decreasing availability for a given zip code.

Data Profile:

The data originally had 83 columns and 455 rows. After consistency checks and cleaning it has 81 columns and 455 rows.

There were no duplicates, but I removed the columns for "location_fips_code", because it didn't seem to have any value I know of, and "population" because it's not needed for my analysis.

Columns	Description	Time Variant / Invariant	Data Type
Location_name	The name of the location	Invariant	Qualitative
location_type	This indicates whether the location is a city, state, metro, or county	Invariant	Qualitative

state	The state of each location	Invariant	Qualitative
county	The county of each location (if applicable)	Invariant	Qualitative
metro	The metropolitan area of each location (if applicable)	Invariant	Qualitative
2017_01	The vacancy percentage of units listed on Apartment List, which meets certain criteria explained above	Variant	Quantitative
One more for each month until	There are over 70 of those columns	Variant	Quantitative
2023_04	The vacancy percentage of units listed on Apartment List, which meets certain criteria explained above	Variant	Quantitative

Limitations:

Each location listed has very thorough historical information. However, there are so few locations listed out of the possible number of different zip codes, that this may not be very helpful for my analysis, if I can't easily match up enough of these locations with the places I'm trying to analyze. Also, as this just mentions vacancy percentage (without actual vacancy numbers), it may be hard to translate this into information worth using, since I'm trying to find out how many apartments are available within a certain price range for each location.

Also, it's possible that something inherent in the methodology of this calculation could be off, or maybe just inaccurate in certain locations. However, with the very narrow scope of this list, that seems unlikely.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's presented publicly and all the information is aggregated without any possible personally identifying information.

4. Apartment List Rent Estimates

Data Source:

This open source data set is publicly available. It was downloaded from apartmentlist.com on their Data & Rent Estimates page: https://www.apartmentlist.com/research/category/data-rent-estimates The "Historic Rent Estimates (Jan 2017 - Present)" was chosen.

Data Collection:

According to Apartment List, the data is calculated "using fully-representative median rent statistics for recent movers taken from the Census Bureau's American Community Survey, extrapolated forward to the current month using a growth rate calculated from real-time lease transactions that take place on our platform. We use a same-unit, repeat-transaction analysis similar to Case-Shiller's approach, comparing only units that are available across both time periods to provide an accurate picture of rent growth. Our approach also corrects for the sample bias inherent in private listing sources to produce results that are representative of the entire rental market."

Data Contents:

This data includes monthly historical rental property rates for over 1000 locations in the US. It includes rates based on "1br", "2br", and "overall" for each location, as well as the population of each location, and the state, country, or metro area (where applicable).

Why this Data:

This data was chosen to help better align trends in anticipated rent costs depending on the size of the unit (for those locations which can be matched here from the food production facilities zip code list). This can also be compared with the Zillow "ZORI" rental rates data to see how close they are, and to see if one or the other is closer to the observed data from currently available rental properties scraped from realtor.com. This can potentially help predict costs for available rental units in the targeted zip code areas.

Data Profile:

The data originally had 84 columns and 3378 rows. After consistency checks and cleaning it has 82 columns and 3378 rows.

There were no duplicates, but I removed the columns for "location_fips_code", because it didn't seem to have any value I know of, and "population" because it's not needed for my analysis.

Columns	Description	Time Variant / Invariant	Data Type
location_name	The name of the location	Invariant	Qualitative
location_type	This indicates whether the location is a city, state, metro, or county	Invariant	Qualitative
state	The state of each location	Invariant	Qualitative
county	The county of each location (if applicable)	Invariant	Qualitative
metro	The metropolitan area of each location (if applicable)	Invariant	Qualitative
bed_size	Whether the unit listed is 1 bedroom or 2 bedrooms, or the overall rent estimate for all unit types	Invariant	Qualitative
2017_01	The vacancy percentage of units listed on Apartment List, which meets certain criteria explained above	Variant	Quantitative
One more for each month until	There are over 70 of those columns	Variant	Quantitative
2023_04	The vacancy percentage of units listed on Apartment List, which meets certain criteria explained above	Variant	Quantitative

Limitations:

Most locations listed have really thorough historical information, but less so (more than 5% missing) when before June 2017 . However, these locations are not the entire US, but rather a certain selection of them for which Apartment List had sufficient information. So this may not be very helpful for my analysis if I can't easily match up enough of these locations with the places I'm trying to analyze. However, it may be helpful to match up these with Zillow's ZORI figure, and if the numbers prove close enough (or if they correlate well), it may help verify the accuracy of Zillow's method, which would indicate that I can better use Zillow's expanded list of zip codes for which they have rent estimates.

Also, it's possible that something inherent in the methodology of this calculation could be off, or maybe just inaccurate in certain locations.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's presented publicly and all the information is aggregated without any possible personally identifying information.

5. Zip code list acquired from major protein company facility locations

Data Source:

This open source data set is publicly available. A google search for "tyson facility address list" returned the data set from this link as the first result: https://www.tysonfoods.com/sites/default/files/2018-07/Tyson%20Foods%20Major%20Locations.pdf The list was compared to an internal list available to Tyson Foods employees and found to be accurate with only minor adjustments needed to be made due to recent facility list changes (i.e. 2 plants closed down, and several plants built - all of which can easily be found in recent news articles).

Data Collection:

This data is published by Tyson Foods as a list of all of their major food production facilities.

Data Contents:

This data includes state, city, zip, and location name for all of Tyson's major food production facilities.

Why this Data:

This data was chosen to narrow the focus of this project to specific areas where a major food producer employs large numbers of workers at its facilities.

Data Profile:

The data originally had 5 columns and 113 rows. After consistency checks and cleaning it has 5 columns and 111 rows. Another version of this dataset was created to include only 1 column (zip codes) and 95 rows - so that it could be imported into my webscraper without any duplicate zip codes.

There were no duplicates in the first list and no columns that I wanted to remove. However, I removed the rows for 2 plants that were recently closed. Columns were renamed to be shorter and more accurate. A couple zip codes listed had the second 4 digit code and that was removed for consistency.

Columns	Description	Time Variant / Invariant	Data Type
location	The name of the food processing facility	Invariant	Qualitative
address	The street address of the facility	Invariant	Qualitative
city	The city of the facility	Invariant	Qualitative
state	The state of the facility	Invariant	Qualitative

zip	The zip code of the facility	Invariant	Qualitative

Limitations:

This data serves a narrow purpose (to identify locations of food processing facilities from a major food processing company), and I don't see any limitations in it being used for that purpose. The only possible limit would be if the data were wrong / inaccurate, possibly not listing some locations or having wrong addresses. Addresses can be easily verified for a list this small and if there were non-public locations, that wouldn't necessarily affect the purposes of this analysis.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's easily searchable on the internet.

6. Realtor.com scraped homes for sale listings for zip codes in list

Data Source:

This data set was compiled by scraping publicly available data from realtor.com on May 5th 2023. This data set is made up of all pages of "for sale" listings for each zip code in the food production facilities zip code list.

Data Collection:

This data was collected by creating a python script to search "for sale" listings on realtor.com for each zip code in the food production facilities zip code list, to then parse the information from each page of listings for each zip code, to select desired data, to create a data set in python, and then export it to a csv. Some zip codes did not have active listings available on the day when the data was scraped and realtor.com automatically returned any listings available within 3 miles of the zip code. These listings will be matched to the desired food production facility zip code during the data wrangling stage of this project.

Data Contents:

This data includes individual "for sale" listings per zip code entered into realtor.com. For each listing it includes the address, city, state, and zip code, the availability ("for sale", "pending", "contingent", or "foreclosure"), the price, the number of beds, baths, and sqft.

Why this Data:

This data was specifically scraped because it was the only way to determine how many properties in each location are below a certain price point, and what those affordable properties look like in terms of sqft, etc. No historical data gave a breakdown of properties available at various price points, so only scraping current data was able to give this information. It is hoped that by comparing this data in an aggregated fashion (i.e. number of listings available and how many of them are below a certain price point), it might be extrapolated together with the other data sets how many homes below a certain price point have historically been available. And the hope is that in using that calculation, that prediction of future availability of affordable housing may also be determined. At the very least, this data can be scraped repeatedly using the same script and up-to-date information about available affordable housing per food production facility location can be given.

Data Profile:

The data originally had 8 columns and 9248 rows. After consistency checks and cleaning it has 11 columns and 6167 rows.

Because this was gathered by a web scraper, some formatting had to be cleaned from the way some listings were displayed on the website. There were a couple hundred duplicates which happened when the webscraper had to be restarted for certain zip codes. There were also a few thousand listings that didn't include square feet information and they were found to be "lots for sale", which is unnecessary for the purpose of this analysis, so those rows were dropped. About 7 listings also didn't have addresses and were dropped. Several listings were posted twice with two different prices and those duplicates were dropped. A few listings were given for selling entire apartment complexes or other large multi-unit buildings (which is beyond the scope of this analysis), so they were dropped. 3 new columns were derived from the "Address2" column - "City", "State", and "Zip", which may aid in linking to some of the other data sets.

Columns	Description	Time Variant / Invariant	Data Type
TysonZip	The zip code of the food processing facility that was put into the URL of the realtor.com web scraper	Invariant	Qualitative
Address1	The full address of a home listed for sale	Invariant	Qualitative
Address2	The city, state, and zip code of a home listed for sale	Invariant	Qualitative
Availability	The availability of the property at the time it was scraped: For Sale, Pending, Contingent, Foreclosure, or Coming Soon	Invariant	Qualitative
Price	The listed price of the property for sale	Invariant (because even though prices sometimes change, that info won't be captured here)	Quantitative
Beds	The number of bedrooms of the property for sale	Invariant	Qualitative
Baths	The number of bathrooms of the property for sale	Invariant	Qualitative
Sqft	The number of square feet of the property for sale	Invariant	Quantitative
City	The city address of the property listed for sale	Invariant	Qualitative
State	The state address of the property listed for sale	Invariant	Qualitative
Zip	The zip code of the property listed for sale	Invariant	Qualitative

Limitations:

This data is scraped directly from the website, so the limitations should only be the limitations of the data publicly listed on realtor.com. However, there may be errors in the webscraper which have not yet been identified and which might cause errors in the data it gathers. Possibly it's communicating in some way with the website that might cause the website to return incorrect or incomplete information. Multiple checks of the data returned indicate that this isn't the case, but it's hard to be sure without doing a 100% manual check (and that would take too long).

Another limitation of this data is that when for sale listings weren't found for certain zip codes, the website automatically provided a list of for sale properties within 3 miles of the zip code listed. This variability might cause challenges for accurately linking it with the other data sets in this project.

Another limitation of this data is that even though the web scraping script is intended to be repeated as often as desired, it is likely that at some point in the future the website will change the way it's coded and/or it will improve it's anti-bot capabilities, and that would break the functionality of the webscraper for gathering future data.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's all information publicly available on public websites (that don't require sign-in). The specific data gathered does include specific addresses, but without any reference to any individual that can be found through this dataset gathered. One possible ethical consideration is that the "terms of use" of realtor.com probably says something about web scrapers / bots not being authorized, but I didn't read that and I wasn't required to agree to a "terms of use" before accessing their website, so any such statement would be non-binding. I believe that US legal courts have also ruled that this type of webscraping is perfectly legal.

7. Realtor.com scraped rental listings for zip codes in list

Data Source:

This data set was compiled by scraping publicly available data from realtor.com on May 5th 2023. This data set is made up of all pages of "for rent" listings for each zip code in the food production facilities zip code list.

Data Collection:

This data was collected by creating a python script to search "for rent" listings on realtor.com for each zip code in the food production facilities zip code list, to then parse the information from each page of listings for each zip code, to select desired data, to create a data set in python, and then export it to a csv. Some rental properties listed a range of unit sizes and unit prices, and in those cases the data was split to create a single listing for the high end of the range and another single listing for the low end of the range. Some zip codes did not have active listings available on the day when the data was scraped and realtor.com automatically returned any listings available within 3 miles of the zip code. These listings will be matched to the desired food production facility zip code during the data wrangling stage of this project.

Data Contents:

This data includes individual "for sale" listings per zip code entered into realtor.com. For each listing it includes the address, city, state, and zip code, the style ("apartment", "condo", "townhome", or "house"), the price, the number of beds, baths, and sqft.

Why this Data:

This data was specifically scraped because it was the only way to determine how many rental properties are available in each location as well as how many are below a certain price point, and what those affordable properties look like in terms of sqft, etc. No historical data gave a breakdown of rental properties available at various price points, so only scraping current data was able to give this information. It is hoped that by comparing this data in an aggregated fashion (i.e. number of rental listings available and how many of them are below a certain price point), it might be extrapolated together with the other data sets how many rental units below a certain price point have historically been available. And the hope is that in using that calculation, that prediction of future availability of affordable rental housing may also be determined. At the very least, this data can be scraped repeatedly using the same script and up-to-date information about available affordable rental housing per food production facility location can be given.

Data Profile:

The data originally had 8 columns and 1610 rows. After consistency checks and cleaning it has 11 columns and 1599 rows.

Because this was gathered by a web scraper, some formatting had to be cleaned from the way some listings were displayed on the website. This included the fact that many rental listings would give a price range and range of unit sizes. When this happened, it was split into two different rows of data, one from the highest part of the range mentioned, and one from the lowest part of the range mentioned. Sometimes a single price was given but with a range of sizes, and in these instances, the smallest of the sizes was kept with the assumption that the "starting price" or "lowest price" was listed (but these rows were not turned into two separate rows). Several rows were dropped because the price wasn't mentioned. 3 new columns were derived from the "Address2" column - "City", "State", and "Zip", which may aid in linking to some of the other data sets.

Columns	Description	Time Variant / Invariant	Data Type
TysonZip	The zip code of the food processing facility that was put into the URL of the realtor.com web scraper	Invariant	Qualitative
Address1	The full address of a unit listed for rent	Invariant	Qualitative
Address2	The city, state, and zip code of a unit listed for rent	Invariant	Qualitative
Style	The type of unit listed for rent: Apartment, House, Townhome, Condo, Other, Condo/Townhome, or Duplex/Triplex	Invariant	Qualitative
Price	The listed price of the unit listed for rent	Invariant (because even though prices of a unit sometimes change over time, that info won't be captured here nor analzyed)	Quantitative
Beds	The number of bedrooms of the unit	Invariant	Qualitative

	listed for rent		
Baths	The number of bathrooms of the unit listed for rent	Invariant	Qualitative
Sqft	The number of square feet of the unit listed for rent	Invariant	Quantitative
City	The city address of the unit listed for rent	Invariant	Qualitative
State	The state address of the unit listed for rent	Invariant	Qualitative
Zip	The zip code of the unit listed for rent	Invariant	Qualitative

Limitations:

This data is scraped directly from the website, so the limitations should only be the limitations of the data publicly listed on realtor.com. However, there may be errors in the webscraper which have not yet been identified and which might cause errors in the data it gathers. Possibly it's communicating in some way with the website that might cause the website to return incorrect or incomplete information. Multiple checks of the data returned indicate that this isn't the case, but it's hard to be sure without doing a 100% manual check (and that would take too long).

Another limitation of this data is that not all units for rent in a given area are listed through Realtor.com. Realtor.com is more likely to not display all units for rent than they are to not display all homes for sale. This is evidenced in the fact that quite a lot of "TysonZips" were showing no rental units available - neither at the actual zip code, nor in a 3 mile radius. And when rental listings weren't found for certain zip codes, the website automatically provided a list of rental units within 3 miles of the zip code listed. This variability might cause challenges for accurately linking it with the other data sets in this project.

Another limitation of this data is that even though the web scraping script is intended to be repeated as often as desired, it is likely that at some point in the future the website will change the way it's coded and/or it will improve it's anti-bot capabilities, and that would break the functionality of the webscraper for gathering future data.

Ethics:

I don't believe there are any ethical challenges with using this data, since it's all information publicly available on public websites (that don't require sign-in). The specific data gathered does include specific addresses, but without any reference to any individual that can be found through this dataset gathered. One possible ethical consideration is that the "terms of use" of realtor.com probably says something about web scrapers / bots not being authorized, but I didn't read that and I wasn't required to agree to a "terms of use" before accessing their website, so any such statement would be non-binding. I believe that US legal courts have also ruled that this type of webscraping is perfectly legal.

Questions to Explore:

The goal of this project is to determine housing affordability in locations with major food processing facilities.

For that reason we want to look at availability of houses for sale and units for rent in each of those locations. We also want to look at the price range of those homes and how many

homes are available within certain price ranges that would typically be affordable to someone working at one of these food processing facilities.

We will also explore the historical trend of houses for sale, number of homes listed, and the average costs of those homes and see what kind of trends might be forecast into the future. This information will be compared with present disaggregated data to try to guess how many homes were available within a certain price range in the past and use that to forecast what expectations might be for the future.

The same efforts will be taken in regards to rental home availability, though this may be harder with only “vacancy rate” information and not number of units available. This is further compounded by the inability to find any historical rental data that is complete for every zip code in the US.

Another question to explore would be to see if housing / rental prices go up or down when units for sale and vacancy percentages increase or decrease.