Zillow Data Organization Project

Trent Seigfried | Apr 8, 2025

Your goal with this project is to create a summary CSV file of data points from Zillow.com.

The raw data source files from Zillow, which I will point to below, are organized such that the historical data is stored in the same file as columns with the date of that data point as the header. In most cases below, **we will want data from the most recent column – the rightmost one.** In the future, when we run this script, we always want to be obtaining data from the most recent column.

You should create a Python script that produces a CSV file with the following columns. Call it Zillow_Data.csv

key_row - This should be column A from this file, for each value for which there is an entry in column C -> Keys. If column C is empty, skip it. You will want to also retain column C (RegionName) and column F (region_type).

regiontype - The value from column F in Keys

regionname - The value from column C in Keys

statename - If regiontype is country, this should be US
If regiontype is state, this should be the two letter postal abbreviation of that state.
If regiontype is metro, this should be the two letter postal abbreviation from regionname.

date - This data is pulled from the following file at Zillow:
https://files.zillowstatic.com/research/public_csvs/zhvf_growth/Metro_zhvf_growth_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

The RegionName in this file should match with most of the regionname values above. You'll want to pull in the value in the BaseDate column. Date should be in MMMM D, YYYY format.

forecastyoypctchange - Using the same file you used for date, pull in the rightmost column instead.

home_value - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv - and pull in the rightmost column.

For home_value for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

home_value_rounded - The value in home_value, rounded to the nearest 1,000

home_value_change_mm - The percent difference between home_value (the rightmost column in the two files for that value) and the value in the *next-to-rightmost* column from those files. For example, when I look at them, the rightmost two columns have headers of 2025-01-31 and 2025-02-28. You'd just get the percent change from the value in 2025-01-31 to 2025-02-28. Remember, this should be *dynamic* and not based on specific columns.

home_value_change_yy - The percent difference between home_value (the rightmost column in the two files for that value) and the value in the *thirteenth-to-rightmost* column from those files. For example, when I look at them, these two columns have headers of 2024-02-29 and 2025-02-28. You'd just get the percent change from the value in 2024-02-29 to 2025-02-28. Remember, this should be *dynamic.*

2016_median_home_value - The median value of columns with a header starting with 2016 in the files from the home_value section.

2017_median_home_value - The median value of columns with a header starting with 2017 in the files from the home_value section.

2018_median_home_value - The median value of columns with a header starting with 2018 in the files from the home_value section.

2019_median_home_value - The median value of columns with a header starting with 2019 in the files from the home_value section.

2020_median_home_value - The median value of columns with a header starting with 2020 in the files from the home_value section.

2021_median_home_value - The median value of columns with a header starting with 2021 in the files from the home_value section.

2022_median_home_value - The median value of columns with a header starting with 2022 in the files from the home_value section.

2023_median_home_value - The median value of columns with a header starting with 2023 in the files from the home_value section.

2024_median_home_value - The median value of columns with a header starting with 2024 in the files from the home_value section.

2025_median_home_value - The median value of columns with a header starting with 2025 in the files from the home_value section.

top_tier - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_uc_sfrcondo_tier_0.67_1.0_sm_sa_month.csv - and pull in the rightmost column.

For top_tier for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_uc_sfrcondo_tier_0.67_1.0_sm_sa_month.csv

bottom_tier - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_uc_sfrcondo_tier_0.0_0.33_sm_sa_month.csv - and pull in the rightmost column.

For top_tier for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_uc_sfrcondo_tier_0.0_0.33_sm_sa_month.csv

sale_price - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/median_sale_price/Metro_median_sale_price_uc_sfr_month.csv

listing_price - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/mlp/Metro_mlp_uc_sfrcondo_sm_month.csv

current_days_to_pending - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/med_doz_pending/Metro_med_doz_pending_uc_sfrcondo_sm_month.csv

dtp_month - The date of the rightmost column in the days_to_pending file in MMMM YY format

days_to_pending - Match the average of the rightmost 12 columns from the current_days_to_pending file, based on regionname above

inventory - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/invt_fs/Metro_invt_fs_uc_sfrcondo_sm_month.csv

Zestimate Data

The next several fields should be scraped from https://www.zillow.com/z/zestimate/ using BeautifulSoup or Selenium.

zestimate_accuracy - For all locations, it's the last updated date on this page - https://www.zillow.com/z/zestimate/. If you cannot get a date, substitute the first date of the current month. Date should be in MMMM D, YYYY format.

For the next few, you'll use the Active Listings table at the top for Metros, States, and Nationwide.

zillow_on_market_median_error - The median error value from the Active Listing Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_on_market_properties - The Homes with Zestimates value from the Active Listing Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_on_market_within_5 - The Within 5% of Sales Price value from the Active Listing Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_on_market_within_10 - The Within 10% of Sales Price value from the Active Listing Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_on_market_within_20 - The Within 20% of Sales Price value from the Active Listing Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

For the next few, you'll use the Off Market table further down the page for Metros, States, and Nationwide.

zillow_off_market_median_error - The median error value from the Off Market Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_off_market_properties - The Homes with Zestimates value from the Off Market Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_off_market_within_5 - The Within 5% of Sales Price value from the Off Market Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_off_market_within_10 - The Within 10% of Sales Price value from the Off Market Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

zillow_off_market_within_20 - The Within 20% of Sales Price value from the Off Market Zestimate tables. Note that although you should be able to get a national match and state matches, the number of metro matches will be limited.

Additional Fields

home_value_one_bed - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_bdrmcnt_1_ uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv - and pull in the rightmost column.

For home_value for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_bdrmcnt_1_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

home_value_two_bed - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_bdrmcnt_2_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv - and pull in the rightmost column.

For home_value for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_bdrmcnt_2_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

home_value_three_bed - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_bdrmcnt_3_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv - and pull in the rightmost column.

For home_value for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_bdrmcnt_3_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

home_value_four_bed - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_bdrmcnt_4_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv - and pull in the rightmost column.

For home_value for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_bdrmcnt_4_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

home_value_five_bed - Match this file based on regionname as above - https://files.zillowstatic.com/research/public_csvs/zhvi/Metro_zhvi_bdrmcnt_5_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv - and pull in the rightmost column.

For home_value for key_rows with regiontype state, use this file instead - https://files.zillowstatic.com/research/public_csvs/zhvi/State_zhvi_bdrmcnt_5_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv

How to Handle Missing Data

These data files won't cover every possible point. Figure out how to fill in the missing data points. Here are a few suggestions:

- If a metro value is missing, see if there's a value for the same state and use that instead.
- If a state value is missing, use the average of the metro values in that same state.
- If the national value is missing, use an average of all states

Explain what you did to produce the missing data points.

Turning In Results

Put your Zillow_Data.csv file and the script or notebook you used to generate it in a new repository on Github, and share that repository in the Slack channel.

If you have any questions, ask them in the Slack channel.