Journal of Information Systems Education

Volume 35 | Issue 4 Article 5

12-15-2024

Teaching Case: Cleaning House: A Case Teaching Data Cleaning Using Real-World Zillow Real Estate Data

Cassandra Artman Collier
Southern Illinois University – Edwardsville, cassaco@siue.edu

Recommended Citation
Collier, Cassandra Artman (2024) "Teaching Case: Cleaning House: A Case Teaching Data Cleaning Using Real-World Zillow Real Estate Data," *Journal of Information Systems Education*: Vol. 35 : Iss. 4 , 456-460.
DOI: https://doi.org/10.62273/OQXC8468
Available at: https://aisel.aisnet.org/jise/vol35/iss4/5

Journal of Information Systems Education

Volume 35
Issue 4
Fall 2024

Teaching Case

Cleaning House: A Case Teaching Data Cleaning Using Real-World Zillow Real Estate Data

Cassandra Artman Collier

Recommended Citation: Collier, C. A. (2024). Teaching Case: Cleaning House: A Case Teaching Data Cleaning Using Real-World Zillow Real Estate Data. *Journal of Information Systems Education*, 35(4), 456-460. https://doi.org/10.62273/OQXC8468

Article Link: https://jise.org/Volume35/n4/JISE2024v35n4pp456-460.html

Received: August 25, 2023
First Decision: February 28, 2024
Accepted: June 10, 2024
Published: December 15, 2024

ISSN: 2574-3872 (Online) 1055-3096 (Print)

Journal of Information Systems Education, 35(4), 456-460, Fall 2024
https://doi.org/10.62273/OQXC8468

Teaching Case
Cleaning House: A Case Teaching Data Cleaning Using Real-World Zillow Real Estate Data

Cassandra Artman Collier
Department of Computer Management & Information Systems
Southern Illinois University – Edwardsville
Edwardsville, IL 62025, USA
cassaco@siue.edu

ABSTRACT

When we imagine the work of a data analyst, we often picture meaningful data analysis and beautiful data visualizations. Although that is an exciting part of the job, data analysts actually spend the majority of their time acquiring, cleaning, and preparing data for analysis. This teaching case guides students through some of the most common data cleaning challenges, including handling missing data, reshaping datasets, splitting columns, and profiling data to anticipate data quality concerns. Students will practice these skills in Microsoft Power BI, a current market leader in data analytics, using real-world, publicly available data from the popular United States real-estate platform Zillow. This case would be a good addition to data analytics, data management, or data visualization classes, or in general information systems courses looking to introduce students to the vital activity of data cleaning.

Keywords: Teaching case, Data cleansing, Data literacy, Data acquisition, Data analytics

1. CASE SUMMARY

This case utilizes a real-world dataset provided free to the public by the popular housing and real-estate platform Zillow. Students are led through the steps of acquiring the data and a variety of data cleansing activities: loading data, promoting headers, filtering rows, splitting, renaming, and removing columns, profiling columns and assessing data quality, handling missing data, reshaping the data through unpivoting, changing data types, identifying and dealing with duplicate data, creating hierarchies, and documenting data provenance. This case utilizes the leading data analytics tool Microsoft Power BI, licenses for which are included for free in Microsoft Office 365 subscriptions and thereby freely available for many university programs. By working through this case, students will learn the basics of data cleansing while also engaging in critical thought processes and class discussions on various data cleansing strategies.

2. CASE TEXT

2.1 Introduction

Data rarely exists in a format that is ready to be analyzed and visualized. Data analysts are estimated to spend up to 80% of their time discovering and preparing data (DalleMule & Davenport, 2017) and "dirty" or bad data is estimated to cost the U.S. $3 trillion per year (Redman, 2016). This is true of real-world housing data, which you are tasked with obtaining, preparing, cleaning, wrangling, and analyzing in this project.

For this project, you will utilize data from the popular housing market application Zillow and will prepare, clean, and analyze it with marketing leading data analytics tool Microsoft Power BI. The goals of the project are to convert housing data into a format easily digestible by Power BI (and other leading market tools) and conduct an analysis of the data.

2.2 Zillow – Overview and Data

2.2.1 Introduction to Zillow. Zillow is a real-estate website and app operating in the United States. Launched in 2006, Zillow's website provides information on the U.S. real-estate market, from prices of for-sale and recently sold homes, estimates of what the homes are worth, mortgage estimates, sale history, and rental prices and history (Zillow Group, 2023b). Zillow includes functions for buyers, sellers, renters, landlords, and realtors to conduct a variety of real estate business dealings (Zillow, 2023). Figure 1 shows a screenshot of the Zillow platform.

Although Zillow's primary purpose is for buying and selling houses, some of the 236 million unique monthly users (Wylie, 2023) seem to be using Zillow "just for fun" (Goodwin, 2021) or treating the app as an addition to or replacement of existing social media apps (Bryant, 2019). Social media accounts like TikTok's "Zillow Gone Wild" (zillowgonewild, n.d.) and Instagram's "Cursed Zillow Listings" (Cursed Zillow Listings, n.d.), among others, garner thousands of viewers to consume content on Zillow's most interesting and unusual real estate listings.

2.2.2 Zillow Data. Zillow stores a variety of housing data and makes many of these datasets publicly available for anyone with an Internet connection to access. The one of interest for our purposes is the Zillow Home Value Index (ZHVI). The ZHVI represents "a measure of the typical home value and

market changes across a given region and housing type" which Zillow provides in both a smoothed value adjusted for seasonality and as a raw value (Allison, 2022).

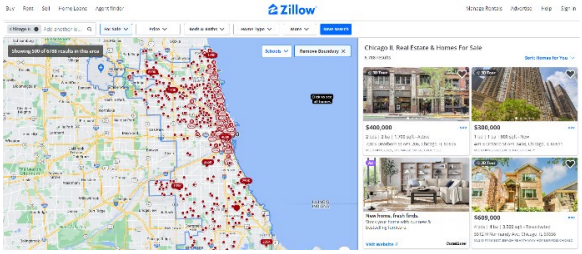


Figure 1. A Screenshot of the Zillow Website, Displaying Homes for Sale in Chicago, IL, in Summer 2023

The ZHVI allows us to address several interesting questions, including:

- How do home prices in different cities and regions of the U.S. compare?
- How have home prices changed over time, in the U.S. as a whole, and in different regions?
- How have housing prices reacted to various events in recent history (the 2008 recession, the COVID-19 pandemic, etc.)?

However, as we will see, the data provided by Zillow (Zillow Group, 2023a) is not in a format to answer these questions using a modern data analytics tool. Next, we will introduce one such tool, and then examine the data cleaning exercises we can undertake to make the data usable for analysis.

2.2.3 Microsoft Power BI. Several modern tools exist for preparing, cleaning, analyzing, and visualizing data. We will focus on Microsoft Power BI, which is a market leading data-analytics tool (Microsoft, 2022). Power BI offers powerful data cleaning, analysis, and visualization functions, and has the additional benefit of being included with Microsoft Office 365 subscriptions (Microsoft, 2022). This means that Power BI licenses are included in an organization's Office 365 subscription, thereby representing cost savings over competing tools. The Microsoft tool also typically has less of a learning curve for new analysts and business users who are likely already familiar with the look and feel of other Microsoft Office products.

This case will focus on the role of Power BI in acquiring, cleaning, and wrangling data to prepare for analysis and visualization. The primary function in Power BI to achieve this goal is the Power Query Editor (Microsoft, 2024b). The Power Query Editor allows users to access extract-transform-load functionality within Power BI and Excel, along with various online Microsoft tools (Microsoft, 2024d). The instructions to follow will guide you in accessing and manipulating the Power Query Editor.

2.2.4 Introduction to Data Cleaning. When we picture data analysts, we likely picture them running complex models, developing beautiful visualizations, and guiding decision-makers with data. However, data analysts actually spend up to 80% of their time discovering and preparing data (DalleMule & Davenport, 2017). This process of data acquisition, cleaning, and wrangling can be incredibly satisfying – like solving a puzzle. More importantly, it is a necessary part of the process for data to be accurately analyzed and attractively visualized. As we will see, even pre-cleaned and published data like that we will acquire from Zillow can still require substantial data cleaning to be useful in an analysis.

2.2.5 Purpose and Business Questions. We are interested in learning how the United States housing market has changed over time, and how it differs in different parts of the country. These insights are valuable to several stakeholders: individuals interested in buying homes, selling homes, or even investing in platforms such as Zillow.

2.2.6 Acquiring Data. To begin, acquire the Zillow data. We will utilize the ZHVI, which Zillow publishes publicly on this website: https://www.zillow.com/research/data/. Visit the website and examine the datasets available. Select the Data Type = ZHVI All Homes (SFR, Condo/Co-op) Time Series, Smoothed, Seasonally Adjusted($) with Geography = Metro & U.S. and download the dataset. The dataset will download as a .csv file. Open it in Excel, rename it to something more meaningful, and save it as a .xlsx file.

2.2.7 Loading Data. Open Power BI and preview the Excel spreadsheet. We can immediately see that several transformations will need to happen with the data before it will be usable, so click "Transform Data" and proceed with the case to begin resolving these issues.

2.2.8 Promoting Headers. One of the major problems we immediately observe upon loading the data is the column names. By default, Power BI does not recognize the headings available and auto-names the columns with unhelpful and unclear names like Column1, Column2, etc. Utilize the "Use First Row as Headers" button in the Power Query Editor to resolve this issue.

2.2.9 Filtering Rows. Now that headers have been promoted, examine the data. What *level of analysis* does the data currently represent?

Upon examining the data, you may note that multiple levels of analysis exist: one row represents country-level data for the entire United States while subsequent rows are based on "Regions" or specific cities within states. Utilizing data all at the same level of analysis will simplify our efforts, so filter out the country-level row by removing rows with RegionType equal to country (the remaining rows have a RegionType of msa or metropolitan statistical areas).

2.2.10 Splitting and Renaming Columns. Next, investigate the RegionName column. Ideally, each column in a dataset would contain one and only one piece of information. This column represents two pieces of information – what are they?

Utilize the split columns feature in Power BI to split this into two columns. Once this operation is completed, you will have two columns: RegionName.1 and RegionName.2. Rename these to something more useful.

2.2.11 Removing Columns. It is good practice to regularly review your dataset for any inconsistencies and redundancies.

After splitting the RegionName column, note that you have created some redundancy in the dataset: there now exist two different columns that contain the same information. Remove one of them.

While you are considering this, examine the remaining columns. Are all useful to your analysis? Hint: anytime a column contains only one value for the entire dataset, it is unlikely to be useful in an analysis using only that dataset.

2.2.12 Profiling Columns and Assessing Data Quality. You have now finished the preliminary steps for adjusting the dataset to something usable. Next, we can begin to evaluate the quality of the dataset, with the use of Power BI's column profiling functionality.

Turn on column profiles, column quality, and column distribution. Maintain the default of column profiling based on the top 1000 rows and answer a few questions:

- What percentage of data is missing from SizeRank?
- What percentage of data is missing from 9/30/2000?
- What is the average price on 9/30/2000?
- What percentage of data is missing from 3/31/2020?
- What is the average price on 3/31/2020?

Consider your answers to these questions. What is your impression of the quality of the data considering these values? Proceed to the next step to evaluate some ways to handle the missing data.

2.2.13 Handling Missing Data. There are several ways to accommodate missing data (or cells with a "null" value) in a dataset. Analysts may elect to filter out rows entirely if they have missing data, they might replace missing values with zeroes, or they might replace missing values with some other value (e.g., the average value for the rest of the dataset), to name a few.

Utilizing the automated feature in Power BI, replace all null values in the 9/30/2000 column with zeroes. Once this operation is completed, re-evaluate – what is the average price for 9/30/2000 now? How does this compare to the average price before the operation?

Repeat these steps for the 3/31/2020 column by replacing all null values in this column with zeroes and re-evaluating the average price on 3/31/2020. How does the magnitude of this change compare with the change for the 9/30/2000 column? Why are they different?

Before proceeding with the case, undo the last two steps (replacing null values with zeroes in the 9/30/2000 and 3/31/2020 columns) by deleting them from the Applied Steps window in the Power Query Editor.

Close and apply changes from the Power Query Editor.

2.2.14 Reshaping the Data: Unpivoting. Having made several basic changes to the data, now attempt to visualize the data. Try building a line chart that visualizes the average house price over time for a given city and state.

You will likely find that, in its current format, the data is very difficult to visualize in this way. To answer the research question of how house prices have changed over time in different regions, we will need to reshape the data through a procedure called *unpivoting*.

Pivoting and unpivoting are both functions by which the same data can be presented in different shapes. You can think of datasets as taking one of two shapes: skinny and long or wide and short (Van Der Merwe, 2019). The process of unpivoting or "flattening" data transforms a dataset from wide and short (many columns and fewer rows) to skinny and long (fewer columns and more rows), which is exactly what we want to do here (Microsoft, 2024c).

To unpivot the data, open the Power Query Editor and edit the query. Select all of the columns labeled as dates and click the Unpivot button. This transforms the dataset so that rather than one column per date that stores a value of the house price on that day, we now have one column that stores the date and one column that stores the price on the date. Rename these two columns to something more appropriate. Close and apply the changes from the Power Query Editor.

Now, reattempt the line chart showing change in home price over time. Now that you have a designated home price column, this should be much easier! However, notice that every date is its own place on the line chart, requiring a lot of scrolling to examine the data (depending on your screen size). It would be more convenient to be able to examine the data at the year level, but this may not yet be possible. We will address this next.

2.2.15 Changing Data Types. Examine the data in the Data view of Power BI or in the Power Query Editor. What is the current data type of your Date column? (*Note: you created and named this column in the unpivot process, so it may be named slightly differently).*

By default, the Date column is not assigned the most ideal data type – which is, of course, Date. Luckily this is an easy change. In the Power Query Editor, select the column and, under the Home tab, find the Data Type dropdown box and change this to Date.

Close and apply your changes and return to your visualization. Now that the column has the correct datatype, note that you have more options with how to visualize this. Select one that you think is appropriate.

2.2.16 Profiling the Data After a Change. Now that you have made a major change to the shape of the data, it is a good idea to conduct another sniff test: a quick, high-level review of raw or processed data or analysis results to evaluate whether they fit with common-sense patterns expected of the data in the given industry or context (Grove, 2019). Evaluate the column profiles and distributions again, both with the top 1,000 records and with the entire dataset (Microsoft, 2024a). Compare your findings with those from earlier – what changes occurred? Why do you think they changed in that direction? Also note the differences between the top 1,000 and the full dataset and consider why those happened.

2.2.17 Identifying and Dealing with Duplicate Data. Create a new page in your Power BI analysis and add a bar chart with City on the x-axis and **sum** of Price on the y-axis. Add City as a filter and turn on Data Labels and then examine: What is the **sum** of price for Columbus?

Next, add State to the x-axis along with City (try rearranging City and State in different orders and see what works best) and re-evaluate: What is misleading or incorrect about the sum of price you found for Columbus earlier? Would you consider this duplicate data? How would you ensure this problem did not occur regularly?

2.2.18 Creating Hierarchies. To make the State-City relationship more usable, one option available to analysts is to create a hierarchy. You may have noticed that the Date column automatically created a hierarchy with the different elements of the date. We can also manually create hierarchies with other kinds of data.

Create a hierarchy to nest city names under states. Try utilizing the hierarchy compared to using both state and city or one or the other. Consider the pros and cons of the two approaches.

2.2.19 Documenting Data Provenance. You have now utilized Power Query Editor to create a dataset that is usable for answering the research questions we were interested in. If desired, go on to create visualizations that can help guide the user to a better understanding of the housing market in the U.S. over time!

A final step in any data cleaning exercise is to ensure that there is proper data provenance for future users and analysts of the project to understand where the data came from and how it has been edited. Open the Power Query Editor and examine the "Analysis Steps" window. This provides some helpful breadcrumbs that an individual could use to get a basic understanding of what has been done with the data, but imagine that you are not at all familiar – it might be challenging to have a full understanding of the changes to the data without additional documentation. Consider what other information would be helpful for an analyst who is new to the data and the project.

2.3 Conclusion

This case has introduced you to a modern real-world dataset, a powerful market-leading data analytics tool, and has walked you through a number of data cleaning activities that are required of data analysts daily. You have practiced many skills that will come in handy in your future work as a data analyst: acquiring, loading, cleaning, wrangling, validating, and analyzing/visualizing data. Congratulations!

3. SUGGESTED RESOURCES AND ADDITIONAL READINGS

Dougherty, J., & Ilyankou, I. (2024). Chapter 4 Clean Up Messy Data in *Hands-On Data Visualization* [Open-Access Web edition]. https://handsondataviz.org/clean.html

Kumar, S. (2021, September 28). *7 Ways to Handle Missing Values in Machine Learning.* Medium. https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

Microsoft (2024). *Clean, Transform, and Load Data in Power BI.* https://learn.microsoft.com/en-us/training/modules/clean-data-power-bi/

Microsoft (2024). *Microsoft Certified: Power BI Data Analyst Associate.* https://learn.microsoft.com/en-us/certifications/power-bi-data-analyst-associate/

Microsoft (2024). *Power Query Documentation*. https://learn.microsoft.com/en-us/power-query/

4. REFERENCES

Allison, M. (2022, February 10). *ZHVI User Guide.* Zillow. https://www.zillow.com/research/zhvi-user-guide/

Bryant, K. (2019, August 29). *Here's an Idea: Replace Your Social Media With Real Estate Apps*. Vanity Fair. https://www.vanityfair.com/style/2019/08/zillow-realtor-realty-app-addiction

Cursed Zillow Listings [@bizarrezillow]. (n.d.). *Real Estate Investment Firm Houses of Horror* [Photos and Videos]. Instagram. https://www.instagram.com/bizarrezillow/

DalleMule, L., & Davenport, T. H. (2017). What's Your Data Strategy? *Harvard Business Review*, 95(3), 112-121.

Goodwin, E. (2021, November 30). *Zillow's Largest User Base Is Browsing Just for Fun*. CivicScience. https://civicscience.com/zillows-largest-user-base-is-browsing-just-for-fun/

Grove, S. (2019, June 22). *5 Pitfalls of Data Science: How the Sniff Test Can Help.* LinkedIn. https://www.linkedin.com/pulse/5-pitfalls-data-science-how-sniff-test-can-help-sally-grove/

Microsoft (2022). *What Is Power BI*. https://powerbi.microsoft.com/en-us/what-is-power-bi/

Microsoft (2024a). *Profile Data to View Statistics (Power Query)*. https://support.microsoft.com/en-us/office/profile-data-to-view-statistics-power-query-79616636-43aa-428f-b14b-f9c5c060f6b2

Microsoft (2024b, September 4). *Query Overview in Power BI Desktop.* https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-query-overview

Microsoft (2024c). *Unpivot Columns (Power Query).* https://support.microsoft.com/en-gb/office/unpivot-columns-power-query-0f7bad4b-9ea1-49c1-9d95-f588221c7098

Microsoft (2024d, January 24). *What Is Power Query?* https://learn.microsoft.com/en-us/power-query/power-query-what-is-power-query

Redman, T. C. (2016, September 22). Bad Data Costs the U.S. $3 Trillion Per Year. *Harvard Business Review*. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

Van Der Merwe, E. (2019, March 20). *Power BI: Pivot and Unpivot Columns.* Data Bear. https://databear.com/power-bi-pivot-and-unpivot-columns/

Wylie, L. (2023, April 28). *Zillow Revenue and Usage Statistics (2023)*. Business of Apps. https://www.businessofapps.com/data/zillow-statistics/

zillowgonewild [@zillowgonewild]. (n.d.). TikTok. https://www.tiktok.com/@zillowgonewild

Zillow Group (2023a). *Housing Data*. https://www.zillow.com/research/data/

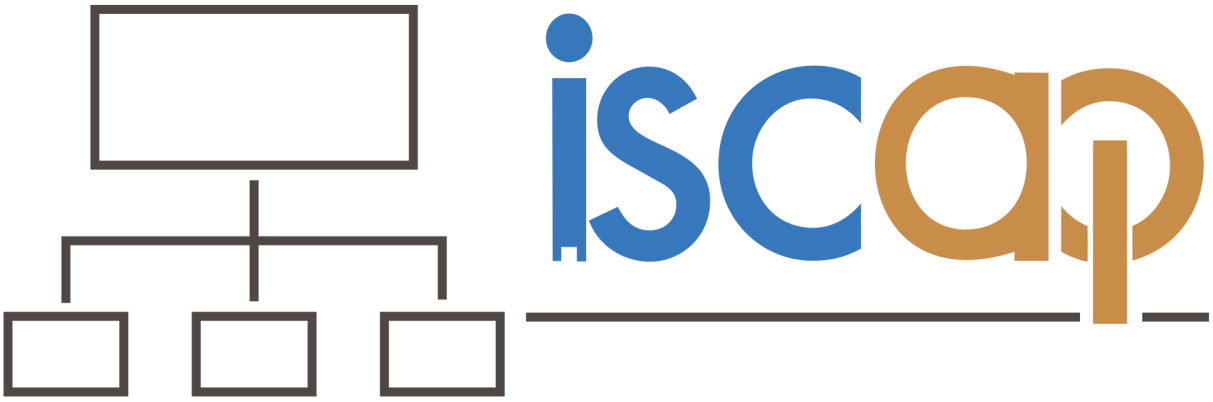
Zillow Group (2023b). Our Services. from https://www.zillowgroup.com/about-us/our-services/

Zillow (2023). *Real Estate, Apartments, Mortgages & Home Values*. https://www.zillow.com/

AUTHOR BIOGRAPHY

Cassandra "Cassie" Collier is an assistant professor in the Computer Management and Information Systems Department at the School of Business, Southern Illinois University Edwardsville. Cassie's research examines IS users' online decision-making and the knowledge and skills needed by today's data analysts. She teaches topics related to data, including big data and data visualization. Prior to her academic career, she worked as a data analyst in a variety of industries, most recently oil and gas.

INFORMATION SYSTEMS & COMPUTING ACADEMIC PROFESSIONALS



STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the ***Journal of Information Systems Education*** have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN: 2574-3872 (Online) 1055-3096 (Print)