Will Millenials be the last generation to afford their own homes?

Analyzing the 2023 housing affordability crisis

Data Management/Analytics Undergraduate Capstone - D195

Mateusz A. Mikulski

Student ID# 001153763

Western Governors University

Table of Contents

[A. Proposal Overview 4](#_Toc2137083979)

[A.1 Research Question or Organizational Need 4](#_Toc1815560889)

[A.2 Context and Background 4](#_Toc1384268585)

[A.3 and A3A Summary of Published Works and Their Relation to the Project 4](#_Toc2079458267)

[Review of Work 1 5](#_Toc694593955)

[Review of Work 2 5](#_Toc475195288)

[Review of Work 3 5](#_Toc1156669356)

[A.4 Summary of Data Analytics Solution 5](#_Toc374341712)

[A.5 Benefits and Support of Decision-Making Process 6](#_Toc1116633891)

[B. Data Analytics Project Plan 6](#_Toc1561262715)

[B.1 Goals, Objectives, and Deliverables 6](#_Toc1217623361)

[B.2 Scope of Project 7](#_Toc427690838)

[B.2.A Included in Project Scope 7](#_Toc1243851001)

[B.2.B Not included in Project Scope 7](#_Toc2036717847)

[B.3 Standard Methodology 7](#_Toc204762054)

[B.4 Timeline and Milestones 8](#_Toc1049840195)

[B.5 Resources and Costs 8](#_Toc313275103)

[B.6 Criteria for Success 9](#_Toc1337049307)

[C. Design of Data Analytics Solution 9](#_Toc764513894)

[C.1 Hypothesis 9](#_Toc1169566000)

[C.2 and C.2.A Analytical Method 9](#_Toc1622461191)

[C.3 Tools and Environments 10](#_Toc53928856)

[C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance 10](#_Toc1386867513)

[C.5 Practical Significance 11](#_Toc157862960)

[C.6 Visual Communication 11](#_Toc1263386613)

[D. Description of Dataset 12](#_Toc34323025)

[D.1 Source of Data 12](#_Toc1867631323)

[D.2 Appropriateness of Dataset 12](#_Toc418252330)

[D.3 Data Collection Methods 12](#_Toc1665164903)

[D.4 Observations on Quality and Completeness of Data 12](#_Toc57902398)

[D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances 13](#_Toc1649406101)

[References 14](#_Toc1322159537)

[Appendix A 14](#_Toc1046046101)

[Title of Appendix 15](#_Toc804972029)

[Appendix B 15](#_Toc128761729)

[Title of Appendix 16](#_Toc1200922991)

[Appendix C 16](#_Toc1897962072)

[Title of Appendix 17](#_Toc1862588384)

[Appendix D 17](#_Toc1292381938)

[Title of Appendix 18](#_Toc390418831)

# A. Proposal Overview

*Are homes currently more unaffordable than ever?*

Amid daily headlines highlighting soaring grocery prices, an economy seemingly reliant on a handful of tech stocks, and a pervasive sense of gloom on social media, the prevailing sentiment suggests that we may be facing unprecedented challenges. As a millennial who purchased their first home during the pandemic, witnessing a 150% surge in housing prices and a 400% spike in interest rates in 2023 is both shocking and deeply concerning. This sentiment is shared among many of my peers, fostering a collective feeling of defeat regarding prospects of homeownership for those who don’t already. Born in 1994, my generation harbors the belief that we might be the last to achieve this milestone, fueled by apprehensions about corporate/private equity firm ownership converting homes into rentals.

Motivated by personal curiosity, my project aims to validate or refute these concerns. I intend to scrutinize various yearly metrics, comparing income statistics, housing prices, interest rates, inflation, average monthly payments, and additional factors such as supply vs. demand, rental percentages, and corporate ownership of single-family homes. My hypothesis posits that, based on correlations among these factors, the average home in 2023 is at its least affordable in the past few decades.

## A.1 Research Question or Organizational Need

Is 2023 the most unaffordable year for buying a home? Headlines ring alarm bells of sky-high home prices, interest rates, inflation, and wages that have not been keeping up with the increased cost of living. That being said, I myself have already lived through one financial crisis in 2008, and my parents through a few of their own. My project aims to find a way to properly formulate affordability, and use a data-driven approach to calculate, graph, and answer this question.

## A.2 Context and Background

As described in my overview, the primary inspiration for this project was of personal curiosity. I also wanted to use my data analytics skills to answer a question I was both passionate about and felt was of significant importance. My approach was to use two Jupyter notebooks: one each for exploratory and explanatory data analysis.

The first step - exploratory analysis, consisted of hunting down, downloading, cleaning, engineering, and exploring reliable datasets pertaining to income, housing prices, and the like, and determining which would be relevant and needed to answer our research question. I performed the cleaning and engineering using Python, and graphed findings with Mathplotlib. I measured positive correlations between datasets using linear regressions and r-values to determine trends, which gave insights on how to proceed.

Once we explored out data and found a reliable way to answer our research question, we pivoted to explanatory data analysis - focusing strictly on eliminating unnecessary data, answering our research question directly, and graphing our results with more detail and clarity.

## A.3 and A3A Summary of Published Works and Their Relation to the Project

### Review of Work 1

As for why housing prices are so high and why it is the worst time to buy a home, articles like this one posit many factors for consideration. It truly is a multi-faceted equation, and Jenni Sisson many reasons one may not initially consider, such as an uncertain economy, high property taces, increased insurance rates, layoffs, and pending legislation such as the YIMBY (Yes, In My BackYard) Act. This inspired me to explore as many datasets as possible in my exploratory data analysis, to determine possible trends and insights.

https://financebuzz.com/worst-time-to-buy-home

### Review of Work 2

My exploratory research led me to this article, which gave me the golden ticket in regards to how to approach answering this question – the Housing Affordability Index. Michael Hyman’s many articles for the National Association of Realtors do a fantastic job of explaining the Housing Affordability Index, and provide insightful graphs and informatics. Without these pages, I wouldn’t know of certain datasets to look for, and how to formulate a mathematical answer to my research question.

https://www.nar.realtor/blogs/economists-outlook/housing-affordability-hits-historical-low-in-august-2023

### Review of Work 3

Lastly, I was able to find this article from 2003 which gives the exact formula we will use for calculating the Housing Affordability Index. https://www.frbsf.org/education/publications/doctor-econ/2003/december/housing-affordability-index/

## A.4 Summary of Data Analytics Solution

**Rubric A4:** The summarized solution is appropriate for the question or need described in part A1, would be able to be realistically implemented, and logically addresses the question or need. The summary is detailed, and *all* of the information in the summary is accurate.

Include a summary of the analytic method(s) and its implementation summarized in B3 and detailed in C4.

## A.5 Benefits and Support of Decision-Making Process

**Rubric A5:** The submission logically explains how the proposed solution will provide a benefit *and* support a decision-making process. The benefits and decision-making process are a realistic consequence of the proposed data analytics solution.

# B. Data Analytics Project Plan

In this part, you will discuss the design details of your capstone data analytics solution.

## B.1 Goals, Objectives, and Deliverables

**Rubric B1:** *All* of the goals, objectives, and deliverables for the project are detailed, realistic, and logically aligned with the project, and they relate to the scope of the project and are aligned with each other.

* Describe each goal (every project should have at least one)
* Describe each objective and how the objectives support the goal(s).
* Describe each deliverable, and how each deliverable supports an objective.

Each goal, objective, and deliverable should align with each other, the project as described in section A, and the timeline provided in section B4. A nested bullet point format will help evaluators identify descriptions and intended alignment.

Example:

* Goal 1: The goal of this project is to …
  + Objective 1.1: Determine if …
    - Deliverable 1.1.1: The deliverable for this objective is …
    - Deliverable 1.1.2: The deliverable for this objective is …
  + Objective 1.2: Provide a …
    - Deliverable 1.2.1: The deliverable for this objective is …

The minimal required goals, objectives, and deliverables are one of each.

## B.2 Scope of Project

**Rubric B2:** The description of the project scope includes what the project will **and will not entail.** The scope details logically align with the goals of the project.

### B.2.A Included in Project Scope

Detail the project scope.

### B.2.B Not included in Project Scope

Detail at least one item outside the project scope.

## B.3 Standard Methodology

**Rubric B3:** The submission logically explains how a specific project planning methodology will be used for the implementation of the proposed project, including specific details on how the methodology will organize the work. The methodology is appropriate and relevant for the implementation of the proposed project.

The methodology is the process you will follow when implementing your solution. Include specific details to adequately describe the steps that will take place in each development phase.

* Identify the standard methodology used.
* Directly connect steps taken in your process to development phases of the standard methodology, e.g., analysis, design, etc.

## B.4 Timeline and Milestones

**Rubric B4:** The provided timeline includes *all* project milestones, including the duration and start and end dates for *each* milestone. *Each* milestone is logically organized and logically sequenced by date, and *each* milestone is realistic and relevant to the project.

Your timeline should align with the deliverables described in section B.1. While a table is not specifically required, it provides a succinct presentation satisfying the requirements of B.4., and it is what evaluators have come to expect.

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Deliverable 1 | 1 day | *Some future date* | *Some future date* |
| Milestone 3 | 36 hours | *Some future date* | *Some future date* |

## B.5 Resources and Costs

**Rubric B5:** The provided list includes *all* necessary resources and *all* associated costs to implement the project. *All* listed resources and costs are realistic and relevant to the proposed project.

Example:

1. Hardware item: $1000
2. Software item 1: No cost
3. Software item 2: No cost
4. 10 work hours: $500 (10 hours at $50 per hour)
5. Item 3: $100

Etc. Include the following: **hardware, software,** and **work hours.** Be realistic as possible when estimating costs. However, this is not a business project and values are not rigorously assessed. The minimum number of listed items is two.

## B.6 Criteria for Success

**Rubric B6:** The submission describes specific criteria for evaluating the success of project execution, and the criteria are measurable and relevant to the proposed project.

Provide a specific metric or criteria for determining success. The criteria or metric should provide specific objective means of assessing success. You should base these criteria on the completion of the analysis methods -not the results. For example, it is acceptable that a test fails to find results statistically significant provided the conclusion and methods are appropriate.

# C. Design of Data Analytics Solution

In this part, you will discuss the design details of your Capstone data analytics solution.

## C.1 Hypothesis

**Rubric C1:** The hypothesis is clearly stated and well aligned with the research question or organizational need identified in part A1.

Provide a hypothesis supporting the research question or organizational need given in section A1. The minimum required hypothesis is one.

## C.2 and C.2.A Analytical Method

**Rubric C2:** The identified analytical method aligns with the proposed solution.

Identify the statistical test or model which will support each hypothesis given in section C.1 and summarize how the method will be performed or developed. The minimum required method is one per hypothesis.

**Rubric C2A:** The submission justifies the chosen analytical method and includes specific, logical reasons why the chosen analytical method is appropriate for addressing the research question or organizational need identified in part A.

For each provided statistical test or model, describe why it is an appropriate choice for supporting the hypothesis (and thus the research question or organizational need from A1).

## C.3 Tools and Environments

**Rubric C3:** The description includes*all* tools and environments used to produce the data analytics solution, and *all* of them are relevant to the project. If third-party code was part of the tools and environment, it has been included.

b

## C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

**Rubric C4:** The submission thoroughly and accurately describes the methods and metrics. The description includes specific details on how the methods and metrics will evaluate statistical significance.

For *each* statistical test, provide the following information:

* A null hypothesis (the opposite of your hypothesis).
* The planned statistical test .
* The metric(s) generated from that test (e.g., a t-stat) from which probability is derived.
* The *alpha* value (denoted α; usually 1% or 5%) that will be used to determine statistical significance (e.g., if α = .05 and *p­*-value = .025 then the null hypothesis will be rejected and there is sufficient evidence to support the hypothesis).

For *each* model, provide the following information:

* The type of model, e.g., supervised regression, supervised classification, etc.
* The algorithm(s) to be used to develop the model.
* The metric(s) to be used to assess performance.
* The benchmark to which the above metric(s) will be compared to determine success of the model(s), e.g., “If the correlation coefficient is ≥ .6, the model will be considered successful…”

**Rubric C4A:** The submission justifies the chosen methods and metrics, including specific, logical, and well-supported reasons for why the chosen methods and metrics are appropriate for the data analytics solution.

For *each* statistical test or model, describe why it is an appropriate choice. This may repeat parts of section C.2.A.

## C.5 Practical Significance

**Rubric C5:** The submission describes how the practical significance of the data analytics solution will be assessed, including specific criteria regarding whether the solution has provided the expected benefits and supported a decision-making process in the context of the chosen research question or organizational need.

Practical significance refers to how meaningful your findings are in practical application. Results are practically significant when the difference is large enough to be meaningful in real life. This is subjective. But at minimum discuss some criteria to judge the practical significance and how this will be used to support the research question or organizational need from A1. Consider including an example of how the client might apply your work discussed in sections C1 through C.4.A.

## C.6 Visual Communication

**Rubric C6:** The submission describes key details about *each* tool and graphical representation that will visually communicate the findings of the data analytics solution, and the described tools and graphical representations will effectively communicate the expected findings.

Task 3, the Project Report, must include graphic visualizations (at least two) for visually communicating elements of your project (see Task 3: G2). Describe a plan to include at least two visualizations of the data, statistical test(s), or model(s). Specifically, name the types of graphs, what they will visualize, and the tools you’ll use to generate the images.

# D. Description of Dataset

## D.1 Source of Data

**Rubric D1:** *Each* source of the data is correctly identified.

Identify each data source. The minimum number of data sources is one.

## D.2 Appropriateness of Dataset

**Rubric D2:** The discussion provides reasons why the dataset is appropriate for the stated goals of the project.

Describe why each data source provided in section D.1 is appropriate for supporting the research question or organizational need from section A.1.

## D.3 Data Collection Methods

**Rubric D3:** The described data collection methods are thorough.

Describe how each data source listed in section D.1 was collected, e.g., “the data was collected by downloading the .csv file from www.kaggle.com/data\_source\_link.html.”

## D.4 Observations on Quality and Completeness of Data

**Rubric D4:** The summary includes logical and accurate observations on *both* the quality and completeness of the data.

Describe both the quality and completeness of the data and any accommodation needed. Often, data is already clean and complete, but it is still necessary to comment on both.

## D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

**Rubric D5:** The discussion accurately addresses the data governance, privacy and security, ethical, legal, and regulatory compliance considerations, and *all* of these relate to the dataset and the proposed project.

Specifically, address how *each* of the following relates to your data and project:

* Data governance
* Privacy
* Security
* Ethical, legal, and regulatory compliance considerations

**Rubric D5A:** *Each* described precaution includes specific details about working with and communicating about the data, and there is a precaution described for *each* of the considerations discussed in part D5. *Each* precaution reasonably manages the risk associated with the considerations discussed in part D5.

Describe any necessary precautions. In cases where an item is not relevant, you must explain why. You only need to discuss measures for handling human data if you collected that data.

# References

Purdue University. (n.d.). Retrieved from APA Formatting and Style Guide (7th Edition) - Purdue OWL: https://owl.purdue.edu/owl/research\_and\_citation/apa\_style/apa\_formatting\_and\_style\_guide/index.html

Scribbr. (2022, December 21). *Free Citation Generator*. Retrieved from Scribbr: https://www.scribbr.com/citation/generator/

Smith, J. (2023). A Generic Journal Article Example. *Generic Journal*, 50-62.

# Appendix A

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix B

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix C

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix D

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.