AI-Driven Urban US Housing Insights for Sustainable City Living

Merve Pakcan Tufenk

January 2025
Research Report
Advanced Analytics for Business Master Program
Advisor: prof. Monica Drăgoicea

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

This study examines the complex interconnections among the United Nations' Sustainable Development Goals (SDGs) through a data-driven analytical approach. By integrating quantitative modeling and structured literature analysis, this research highlights key synergies and trade-offs across economic, social, and environmental dimensions. A systematic literature review, following the PRISMA methodology, ensures methodological rigor in identifying high-impact studies. The findings contribute to a deeper understanding of how progress in specific SDGs influences broader sustainability outcomes, offering insights that enhance strategic planning and interdisciplinary collaboration. This study provides a structured framework for assessing SDG interdependencies, supporting more integrated and evidence-based approaches to sustainable development.

Chapter 1

First Semester

1.1 Introduction

The commitment to achieving sustainable development involves balancing its economic, social, and environmental dimensions in an integrated manner UN, 2015.

At the UN headquarters in New York, the Open Working Group, established by the UN General Assembly, proposed 17 Sustainable Development Goals (SDGs) and 169 targets, along with a preliminary set of 330 indicators in March 2015. Some goals build on the Millennium Development Goals, while others introduce new concepts Hák et al., 2016.

The primary aim of the Sustainable Development Goals (SDGs) is to achieve a better and more sustainable future by 2030. These goals focus on eradicating poverty and hunger, reducing inequalities, fostering inclusive and peaceful societies, promoting human rights and gender equality, and ensuring the protection of the planet and its natural resources for current and future generations UN, 2015.The SDGs aim to drive action on critical global issues, relying on efforts from governments, the private sector, civil society, and individuals. Progress is monitored through voluntary reviews that assess different goals and their interactions, with a particular focus on SDG 17 through annual reporting UN, 2016.Moving towards a more comprehensive and inclusive framework for sustainable development represents a significant step forward. As outlined in the 2030 Agenda for Sustainable Development, the SDGs are designed to be integrated and indivisible, ensuring a balanced approach that considers economic, social, and environmental dimensions UN, 2015.

All SDGs interact with one another – by design they are an integrated set of global priorities and objectives that are fundamentally interdependent Griggs et al., 2017, p.8. Understanding the range of positive and negative interactions among SDGs is key to unlocking their full potential at any scale, as well as to ensuring that progress made in some areas is not made at the expense of progress in others Griggs et al., 2017. This interconnection highlights the necessity of treating the SDGs as a cohesive system rather than isolated goals. Addressing them individually risks overlooking critical synergies and trade-offs that exist across their targets.

Although the SDGs were designed as an integrated system, their implementation often lacks a structured approach to managing goal interdependencies Scharlemann et al., 2020. Positive and negative interactions between targets are frequently overlooked, leading to uncoordinated strategies that may unintentionally weaken progress in related areas Nilsson et al., 2018. Addressing SDGs in isolation risks missing critical synergies and trade-offs, ultimately reducing the effectiveness of sustainability efforts Pradhan et al., 2017. Furthermore, existing policy and evaluation frameworks frequently fail to capture these systemic interactions, resulting in fragmented decision-making Bennich et al., 2020.

The interconnected and complex nature of the SDGs requires a comprehensive approach to measuring progress that considers their overlapping and interdependent dynamics. To effectively evaluate progress in these goals, it is vital to develop specialized Key Performance Indicators (KPIs) that can reflect the multifaceted relationships and interactions inherent in the SDG framework. As Pradhan et al., 2017 highlight, leveraging synergies and addressing trade-offs are crucial to achieving meaningful progress. This paper seeks to examine these interconnections and dependencies, proposing a robust framework of KPIs that ensures a holistic and accurate assessment of sustainability outcomes, ultimately facilitating more informed and integrated decision-making.

The structure of this paper is as follows: Section 2 outlines the research methodology, detailing the PRISMA approach for systematically collecting and selecting relevant articles. Section 3 evaluates existing literature on data-driven approaches to sustainability and SDGs, exploring possible collections, assessing whether information has been used to demonstrate real connections, and analyzing broader implications to understand the possible relationships between them.Section 4 discusses the findings, analyzing their implications for sustainability and exploring interdependencies between data-driven strategies. Finally, Section 5 summarizes the conclusions, reflecting on the study's contributions and suggesting directions for future research.

1.2 Methodology

This study employs a structured approach to systematically collect and analyze relevant academic literature on Sustainable Development Goals (SDGs). The methodology consists of two main phases: article collection and text preparation, ensuring that the selected literature aligns with high-impact scientific standards and methodological rigor.

1.2.1 Article Collection

Search Strategy

To ensure a comprehensive and high-quality literature review, a structured search strategy was implemented using Google Scholar for initial broad coverage and Web of Science for a refined selection of peer-reviewed studies.

The search queries included terms such as:

- "A systematic study of sustainable development goal (SDG) interactions"
- "SDG trade-offs and synergies"
- "Data-driven approaches to SDG measurement"

Filtering and Data Source Validation

- Language Filter: To maintain consistency in interpretation, only English-language publications were included.
- Dataset Use: One dataset from Web of Science was used for validation.
- Time Frame: The research focused on studies published from 2016 onwards.
- Source Types: The majority of the selected sources are peer-reviewed journal articles, ensuring a strong academic foundation. Books, policy reports, and conference papers were considered selectively for their methodological or practical contributions.

Additionally, priority was given to highly cited studies, authors with a strong research impact (H-index), and articles published in high-ranking journals (Q1 and Q2) indexed in Web of Science to ensure academic credibility.

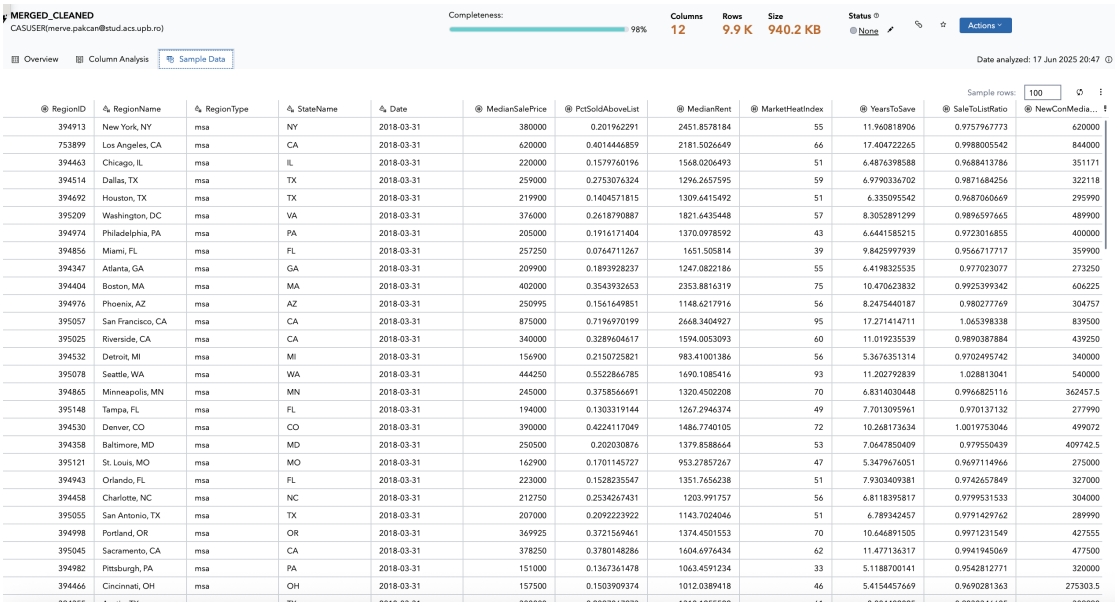
1.2.2 Text Preparation

The literature review process in this study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, a widely recognized framework for conducting systematic literature reviews, ensuring transparency, completeness, and comparability Moher et al., 2009.

The article selection process was conducted in four key stages:

- Identification: A comprehensive literature search was conducted using Web of Science, applying predefined keywords and inclusion criteria.
- Screening: Duplicate records were removed, followed by a title and abstract review to eliminate studies unrelated to SDG interactions and analytics.
- Eligibility: The full texts of the remaining studies were assessed for methodological quality, data availability, and relevance to the research objectives.
- Inclusion: The final set of highly relevant studies was selected for analysis. The selection process prioritized empirical data, novel methodologies, and critical insights into SDG interdependencies.

A PRISMA flow diagram was created to visually represent the selection process, ensuring clarity and reproducibility (Figure 1.1). By adopting PRISMA guidelines, this



MERGED_CLEANED
CASUSER(merve.pakcan@stud.acs.upb.ro)

Completeness: 98%

Columns 12 | Rows 9.9 K | Size 940.2 KB

Status: None

Actions

Overview | Column Analysis | Sample Data

Date analyzed: 17 Jun 2025 20:47

Sample rows: 100

RegionID	RegionName	RegionType	StateName	Date	MedianSalePrice	PctSoldAboveList	MedianRent	MarketHeatIndex	YearsToSave	SaleToListRatio	NewConMedia...
394913	New York, NY	msa	NY	2018-03-31	380000	0.201962291	2451.8578184	55	11.960818906	0.9757967773	620000
753899	Los Angeles, CA	msa	CA	2018-03-31	620000	0.4014446859	2181.5026649	66	17.404722265	0.9988005542	844000
394463	Chicago, IL	msa	IL	2018-03-31	220000	0.1579760196	1568.0206493	51	6.4876398588	0.9688413786	351171
394514	Dallas, TX	msa	TX	2018-03-31	259000	0.2753076324	1296.2657595	59	6.9790336702	0.9871684256	322118
394692	Houston, TX	msa	TX	2018-03-31	219900	0.1404571815	1309.6415492	51	6.335095542	0.9687060669	295990
395209	Washington, DC	msa	VA	2018-03-31	376000	0.2618790887	1821.6435448	57	8.3052891299	0.9896597665	489900
394974	Philadelphia, PA	msa	PA	2018-03-31	205000	0.1916171404	1370.0978592	43	6.6441585215	0.9723016855	400000
394856	Miami, FL	msa	FL	2018-03-31	257250	0.0764711267	1651.505814	39	9.8425997939	0.9566717717	359900
394347	Atlanta, GA	msa	GA	2018-03-31	209900	0.1893928237	1247.0822186	55	6.4198325535	0.977023077	273250
394404	Boston, MA	msa	MA	2018-03-31	402000	0.3543932653	2353.8816319	75	10.470623832	0.9925399342	606225
394976	Phoenix, AZ	msa	AZ	2018-03-31	250995	0.1561649851	1148.6217916	56	8.2475440187	0.980277769	304757
395057	San Francisco, CA	msa	CA	2018-03-31	875000	0.7196970199	2668.3404927	95	17.271414711	1.065398338	839500
395025	Riverside, CA	msa	CA	2018-03-31	340000	0.3289604617	1594.0053093	60	11.019235539	0.9890387884	439250
394532	Detroit, MI	msa	MI	2018-03-31	156900	0.2150725821	983.41001386	56	5.3676351314	0.9702495742	340000
395078	Seattle, WA	msa	WA	2018-03-31	444250	0.5522866785	1690.1085416	93	11.202792839	1.028813041	540000
394865	Minneapolis, MN	msa	MN	2018-03-31	245000	0.3758566691	1320.4502208	70	6.8314030448	0.9966825116	362457.5
395148	Tampa, FL	msa	FL	2018-03-31	194000	0.1303319144	1267.2946374	49	7.7013095961	0.970137132	277990
394530	Denver, CO	msa	CO	2018-03-31	390000	0.4224117049	1486.7740105	72	10.268173634	1.0019753046	499072
394358	Baltimore, MD	msa	MD	2018-03-31	250500	0.202030876	1379.8588664	53	7.0647850409	0.979550439	409742.5
395121	St. Louis, MO	msa	MO	2018-03-31	162900	0.1701145727	953.27857267	47	5.3479676051	0.9697114966	275000
394943	Orlando, FL	msa	FL	2018-03-31	223000	0.1528235547	1351.7656238	51	7.9303409381	0.9742657849	327000
394458	Charlotte, NC	msa	NC	2018-03-31	212750	0.2534267431	1203.991757	56	6.8118395817	0.9799531533	304000
395055	San Antonio, TX	msa	TX	2018-03-31	207000	0.2092223922	1143.7024046	51	6.789342457	0.9791429762	289990
394998	Portland, OR	msa	OR	2018-03-31	369925	0.3721569461	1374.4501553	70	10.646891505	0.9971231549	427555
395045	Sacramento, CA	msa	CA	2018-03-31	378250	0.3780148286	1604.6976434	62	11.477136317	0.9941945069	477500
394982	Pittsburgh, PA	msa	PA	2018-03-31	151000	0.1367361478	1063.4591234	33	5.1188700141	0.9542812771	320000
394466	Cincinnati, OH	msa	OH	2018-03-31	157500	0.1503909374	1012.0389418	46	5.4154457669	0.9690281363	275303.5

Figure 1.1: Systematic literature review and reporting - PRISMA Flow Diagram

study enhances the credibility and methodological transparency of its findings, providing a structured foundation for analyzing SDG trade-offs and synergies.

The study selection followed a structured four-stage approach (illustrated in the PRISMA diagram below):

- Screening: A total of 247 records were retrieved from Web of Science, with 1 duplicate removed. The remaining 246 studies were screened based on title and abstract, leading to the exclusion of 130 studies that did not align with the research objectives.

- Full-Text Review: Out of 116 studies retrieved, 7 were inaccessible due to institutional restrictions or paywalls. The remaining 109 full-text articles underwent a detailed evaluation for methodological rigor, data quality, and relevance.

- Exclusions and Validation: A total of 89 studies were excluded, including 78 unrelated to research topics, 6 with methodological weaknesses (insufficient data or unclear methods), and 5 non-English studies. The eligibility assessment ensured that only high-quality and impactful studies progressed further.

- Final Inclusion and Quality Assurance: A total of 20 studies met all inclusion criteria and were selected for analysis. A final validation step was conducted to ensure consistency and accuracy in paper selection, reducing the risk of bias. The selected studies prioritized empirical data, novel methodologies, and key insights into SDG interdependencies.

1.3 Results: State-of-the-art

To evaluate the state-of-the-art for the topic of this report, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method has been used Moher et al., 2009, Sarkis-Onofre et al., 2021. In this section, the existing literature will be reviewed to investigate and understand the possible relationships among the Sustainable Development Goals (SDGs). By examining current studies, this part aims to identify the synergies and potential trade-offs within the SDG framework, providing a comprehensive understanding of how these goals interact and influence one another.

1.3.1 Synergies Across SDGs

Many SDGs create positive reinforcement loops, where progress in one area accelerates improvements in others. The following key synergies have been highlighted in the literature:

- Social and Economic Synergies: SDG 1 (No Poverty) and SDG 2 (Zero Hunger) significantly contribute to SDG 3 (Good Health and Well-being) by improving nutrition and living conditions Barbier and Burgess, 2019 Pradhan et al., 2017. Enhanced health outcomes, in turn, increase productivity, thereby strengthening SDG 8 (Decent Work and Economic Growth).
- Expanding clean water access (SDG 6) and renewable energy solutions (SDG 7) fosters economic growth, improves public health, and enhances overall quality of life Pradhan et al., 2017. These initiatives create a positive feedback loop, where improved water and energy availability boost agricultural productivity, industrial efficiency, and sustainable urban development. Furthermore, aligning these advancements with responsible consumption and production (SDG 12) enhances their long-term sustainability, ensuring that economic benefits are maintained while preserving natural resources Fuso Nerini et al., 2018.
- Urbanization and Innovation: Sustainable urban development (SDG 11) serves as a catalyst for technological advancements and infrastructure improvements (SDG 9), which in turn foster long-term economic stability and resilience UN, 2019. Well-planned urbanization enhances resource efficiency, reduces inequalities, and stimulates economic opportunities. By integrating smart city planning and sustainable infrastructure, urbanization can drive green innovations, optimize energy consumption, and create more inclusive economic growth, further strengthening synergies among SDGs. Aligning these developments with responsible consumption and production (SDG 12) is essential to ensuring long-term sustainability.

1.3.2 Trade-Offs and Challenges in SDG Implementation

Despite these synergies, some SDGs inherently conflict with others, creating policy dilemmas that necessitate strategic trade-offs. The following key trade-offs have been identified:

- Economic Growth vs. Environmental Protection: While industrialization (SDG 9) and economic growth (SDG 8) contribute to poverty alleviation (SDG 1), they also increase energy consumption and environmental degradation, potentially conflicting with SDG 12 (Responsible Consumption and Production) and SDG 13 (Climate Action) Fuso Nerini et al., 2018. A key challenge is balancing economic growth with sustainable production models.

- Food Security vs. Ecosystem Sustainability: Expanding agricultural production to achieve food security (SDG 2) often results in higher water consumption (SDG 6) and deforestation (SDG 15). Sustainable agricultural practices and biodiversity conservation strategies are crucial to minimizing these environmental trade-offs.

- Climate Action vs. Employment Stability: The transition to a low-carbon economy (SDG 13) may disrupt traditional industries, particularly those reliant on fossil fuels, leading to short-term job losses and economic instability Lusseau and Mancini, 2019. Addressing these employment challenges requires equitable transition policies that support reskilling and job creation in emerging green sectors.

1.3.3 The Role of Integrated Policymaking

The interconnected nature of the SDGs means that progress in one area has cascading effects across multiple goals. Recognizing these interactions allows policymakers to:

- Maximize synergies by aligning economic, social, and environmental objectives.

- Mitigate trade-offs through cross-sector strategies such as sustainable resource management, circular economy models, and just transition policies.

- Leverage global partnerships (SDG 17) to enhance knowledge-sharing, financial support, and policy alignment, ensuring balanced progress across all SDGs Moyer and Bohl, 2019.

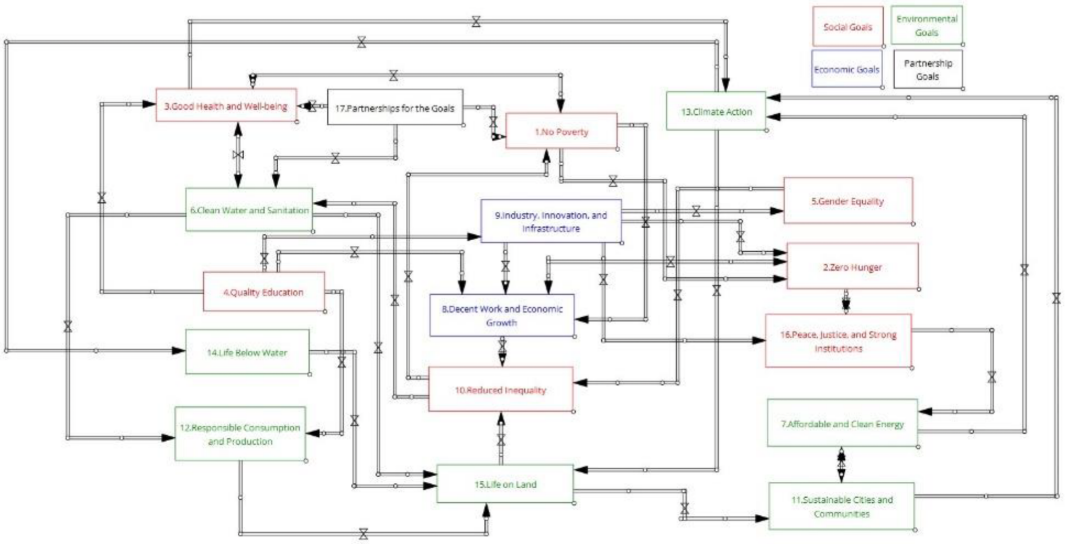


Figure 1.2: Stock-Flow Diagram Illustrating Interdependencies Among SDGs. Source: Luchian et al., 2025

The diagram above illustrates the complex interdependencies between social, economic, and environmental SDGs Luchian et al., 2025. It highlights the reinforcing effects among key goals such as poverty reduction (SDG 1), economic growth (SDG 8), and industrial innovation (SDG 9) while also visualizing the potential conflicts, such as those between climate action (SDG 13) and industrial expansion (SDG 9). Understanding these linkages is essential for designing integrated policies that balance sustainability priorities.

Furthermore, SDG 17 (Partnerships for the Goals) acts as a fundamental connector, linking economic (SDG 8, 9), social (SDG 1, 3, 4), and environmental (SDG 6, 7, 12, 13) dimensions. By fostering global cooperation, knowledge-sharing, and resource mobilization, SDG 17 plays a key role in ensuring a holistic and coordinated approach to achieving sustainable development.

1.4 Discussion

1.4.1 The Role of Data in SDG Implementation and Monitoring

The effective implementation and monitoring of Sustainable Development Goals (SDGs) depend on decision-makers having access to adequate data and robust analytical tools. By leveraging data analytics techniques, valuable insights and patterns can be extracted from comprehensive data sources, enabling informed decision-making to support sustainable growth. However, poor-quality, outdated, or incomplete data significantly hinder the ability to accurately assess SDG progress, particularly when making comparisons

over time or across nations. To achieve meaningful progress, the availability of reliable, consistent, and comparable data is essential. Moreover, the lack of regularly updated and publicly accessible indicators further exacerbates the challenges of tracking SDGs effectively Nilashi et al., 2023.

Achieving sustainable development goals relies heavily on the availability of high-quality data, as poor data quality significantly undermines decision-making and analytical accuracy. The lack of high-quality data is a common pitfall, further complicating efforts to generate reliable insights Meng, 2021.

Achieving SDGs relies heavily on high-quality data, as poor data quality significantly undermines decision-making and analytical accuracy (Meng, 2021). The integration of data analytics plays a pivotal role in addressing these challenges, ensuring data-driven, evidence-based policymaking. However, data accessibility and quality remain critical barriers that affect the assessment and deployment of SDG-related initiatives (Dang et al., 2014). Addressing external pressures, such as privacy and security concerns, and fostering robust governance frameworks are also crucial to ensuring the responsible use of big data for SDG progress Salleh and Janczewski, 2019.

Therefore, adopting a comprehensive and integrated data-driven approach is essential to holistically measure and understand sustainability. This requires synthesizing diverse data sources and aligning them with the interconnected nature of SDGs, ultimately providing a robust framework for sustainability assessment and informed decision-making Teh and Rana, 2023.

1.4.2 Challenges in SDG Data Collection and Availability

Understanding how SDG-related data is collected is essential for evaluating global progress. However, data collection presents significant challenges, as it requires contributions from various national and international sources and the use of innovative statistical methodologies. The UN SDG database integrates data from nearly 200 sources, yet a considerable portion remains missing, particularly for low-income countries. Only about 19 percent of the required data is available for comprehensively tracking global SDG progress Dang and Serajuddin, 2020.

To address these gaps, international organizations play a crucial role in data curation, standardization, and imputation-based statistical techniques, which can serve as cost-effective solutions when actual survey data is unavailable. Moreover, evolving data needs call for the inclusion of alternative indicators, such as subjective well-being measures, which complement traditional economic statistics. Strengthening national statistical capacity and fostering collaboration between stakeholders remain essential steps in improving the availability and reliability of SDG-related data.

1.4.3 Empirical Evidence on SDG Interconnections

As seen in the Figure 1.3, a comprehensive review of methods used to analyze SDG interdependencies is provided with categorizing them into argumentative, literature-based, linguistic, simulation, statistical, and other quantitative approaches Horvath et al., 2022.

The findings highlight that different methods serve distinct purposes—argumentative models like causal loop diagrams conceptualize systemic connections, while simulation techniques such as agent-based modeling enable dynamic scenario testing. Statistical methods, including correlation analysis and regression models, offer empirical insights but often struggle with causality Horvath et al., 2022.

As research shifts towards data-driven methodologies, integrating both qualitative and quantitative approaches has become essential. While previous studies have relied heavily on literature synthesis and conceptual modeling, recent advancements emphasize computational and empirical techniques Horvath et al., 2022. While other studies have used different techniques, this project will utilize a data-driven approach and system dynamics.

1.5 Conclusion

A data-driven approach will be essential for effectively assessing sustainability. This requires integrating diverse data sources and aligning them with the interconnected nature of Sustainable Development Goals (SDGs), ultimately providing a robust framework for sustainability assessment and evidence-based decision-making Dang and Serajuddin, 2020.

Despite challenges in data accessibility, quality, and standardization, studies have demonstrated real connections between SDGs, reinforcing the need for data-driven policymaking. Moving forward, the integration of data analytics and real-time monitoring will be key to improving SDG tracking and ensuring sustainable development pathways.

1.6 Future Research Directions

1.6.1 System Dynamics Method for SDG Analysis

To better understand SDG interdependencies, system dynamics models and stock-flow diagrams will be explored. These tools can help visualize feedback loops, time delays, and trade-offs in sustainability policies. Future research should apply these techniques to analyze interactions among goals like poverty reduction (SDG1), health (SDG3), and economic growth (SDG8), as well as potential conflicts, e.g., between clean energy (SDG7) and responsible consumption (SDG12). Incorporating system dynamics modeling will provide a more comprehensive view of sustainability challenges and support data-driven policy decisions.

Utilizing Large-Scale Databases for SDG Tracking

Enhancing real-time SDG monitoring will require integrating multiple databases from various perspectives. Key sources for future research include:

- Industrial Ecology Data Commons (IEDC): Provides diverse industrial ecology datasets, including resource use, material flows, and input-output data, supporting

Categories and assigned methods to analyse SDG entity interactions.

Category	Method	Refs.
Argumentative	Bayesian belief network (BBN)	(Hall et al., 2018)
	Causal loop diagram (CLD)	(Zhang et al., 2016)
	Cross-impact matrix (CI matrix)	(Allen et al., 2019a; Dawes, 2020; Kumar et al., 2018; Weitz et al., 2018; Zaini and Akhtar, 2019; Zelinka and Amadei, 2017)
	Structured elicitation of expert information (Expert)	(Allen et al., 2019a; Bhaduri et al., 2016; Hall et al., 2018; Hazarika and Jandl, 2019; Jaramillo et al., 2019; Waage et al., 2015; Wieser et al., 2019)
	Nilsson scale (N Scale)	(Allen et al., 2019a; Fader et al., 2018; Fuso Nerini et al., 2019; Hall et al., 2017; Hazarika and Jandl, 2019; Jaramillo et al., 2019; McCollum et al., 2018; Nilsson et al., 2016; Singh et al., 2018; Weitz et al., 2018; Zelinka and Amadei, 2017)
Literature	Non-systematic literature review (Non-syst)	(Alcamo, 2019; Bringezu, 2018; Fisher et al., 2017; Haines et al., 2017; Hazarika and Jandl, 2019; Manandhar et al., 2018; Morton et al., 2017; Pandey and Kumar, 2018; Recuero Virto, 2018; Swamy et al., 2018; Wydra et al., 2019)
	Semi-systematic literature review (Semi-syst)	(Bangert et al., 2017; De Paiva Serôa Da Motta, 2019; Engström et al., 2018; Fuso Nerini et al., 2019, 2018; Hanjra et al., 2016; Hepp et al., 2019; Schroeder et al., 2019)
	Systematic literature review (Syst)	(Alcamo, 2019; Blicharska et al., 2019; Davide et al., 2019)
	Review of case studies (Case studies)	(Alcamo, 2019; Velis et al., 2017)
Linguistic	Keyword analysis (KWA)	(De Paiva Serôa Da Motta, 2019; Le Blanc, 2015; Nugent et al., 2018)
Simulation	Agent based modelling (ABM)	(Wang et al., 2019)
	Computable general equilibrium models (CGE)	(Banerjee et al., 2019; Campagnolo and Davide, 2019; Doelman et al., 2019; Lucas et al., 2019; Matsumoto et al., 2019; Schütze et al., 2017)
	Energy system models (ESM)	(Engström et al., 2019; Vandyck et al., 2018)
	Integrated assessment models (IAM)	(Doelman et al., 2019; Fujimori et al., 2019; Gao and Bryan, 2017; Heck et al., 2018; Hutton et al., 2018; Lucas et al., 2019; Matsumoto et al., 2019; Obersteiner et al., 2016; Rao et al., 2016; von Stechow et al., 2016; Zhang et al., 2019)
	System dynamics modelling (SD)	(Allen et al., 2019b; Collste et al., 2017; Dawes, 2020; Pedercini et al., 2019, 2018; Spaiser et al., 2017)
Other quantitative	Accounting framework (Account)	(Engström et al., 2018)
	Network analysis (NWA)	(Allen et al., 2019a; Dörgő et al., 2018; Feng et al., 2019; Jaramillo et al., 2019; Jiménez-Aceituno et al., 2020; Kunčič, 2019; Le Blanc, 2015; Lim et al., 2018; Lusseau and Mancini, 2019; Mainali et al., 2018; McGowan et al., 2019; Nugent et al., 2018; Sebestyén et al., 2019a, 2019b; Weitz et al., 2018; Zelinka and Amadei, 2017)
	Environmentally-extended multi-regional input-output models (IO)	(Hubacek et al., 2017; Scherer et al., 2018)
Statistical	Advanced sustainability analysis (ASA)	(Mainali et al., 2018)
	Autoregressive distributive lag bounds test (ARDL)	(Ngarava et al., 2019)
	Correlation analysis (Corr)	(Brecha, 2019; Donaires et al., 2019; Kroll et al., 2019; Mainali et al., 2018; Ngarava et al., 2019; Pradhan et al., 2017; Sebestyén et al., 2019a, 2019b)
	Cox proportional hazards models (CPH)	(Akinyemi et al., 2018)
	Descriptive statistics (Descr)	(Howden-Chapman et al., 2020)
	Generalised method of moments (GMM)	(Matthew et al., 2019; Shahbaz et al., 2019)
	Joint correspondence analysis (JCA)	(Ulman et al., 2018)
	Linear mixed effect models (LMM)	(Lusseau and Mancini, 2019)
	Pairwise granger causality test (PGC)	(Ngarava et al., 2019)
	Principal component analysis and Factor analysis (PCA&FA)	(Donaires et al., 2019; Feng et al., 2019; Sen and Ongsakul, 2018; Spaiser et al., 2017)
	Quantile regression, bootstrapped (Q Reg)	(Sinha et al., 2020)
	Regression analysis (Reg)	(Buonocore et al., 2019; Cluver et al., 2016; Hall et al., 2017; Malerba, 2019; Obersteiner et al., 2016; Ramos et al., 2018; Ulman et al., 2018)

Figure 1.3: Source: Horvath et al., 2022

sustainability assessments (https://www.database.industrialecology.uni-freiburg.de).

- GDELT (Global Database of Events, Language, and Tone): Tracks social, economic, and political events affecting SDG progress (https://www.gdeltproject.org/).
- The Proxy Indicator Approach will be explored to assess SDG progress when direct measurements are unavailable. Using indirect indicators as proxies can enhance sustainability assessments by providing alternative insights into complex systems.

Combining these databases with predictive analytics will enable more dynamic and precise sustainability assessments, supporting data-driven evaluations.

1.7 Final Remarks

Future research will prioritize the integration of data-driven methodologies with dynamic modeling to improve sustainability tracking and comprehensive assessments. While synergies between SDGs create significant opportunities for integrated progress, trade-offs remain a critical challenge. Addressing these contradictions requires multi-sectoral, data-driven strategies that incorporate diverse socio-economic contexts. Future research should focus on refining analytical frameworks that balance economic growth, environmental sustainability, and social equity, ensuring that sustainability efforts are both effective and inclusive UN, 2019.

By leveraging advanced data analytics, system dynamics approaches, proxy indicator, and stock-flow diagram, SDG progress can be more effectively monitored, leading to deeper insights and enhanced sustainability evaluations.

Chapter 2

Second Semester

2.1 Introduction

2.1.1 Narrowing the Scope: From SDGs to Urban Housing

In the previous semester, my research focused on the broader landscape of sustainable development, exploring the interconnected nature of the Sustainable Development Goals (SDGs) through a data-driven framework. While this approach revealed important systemic relationships, its wide scope posed certain limitations in terms of practical application and contextual depth.

To address this, the focus of this semester's project was narrowed to a specific and highly relevant domain analyzing of sustainable housing market in USA. Housing operates at the intersection of economic, social, and environmental dimensions, making it a valuable lens for examining sustainability in a localized and measurable way. It not only reflects broader systemic challenges such as inequality, affordability, and urban planning but also directly impacts the livability and resilience of cities.

While the first semester concentrated on exploring SDGs from a global perspective through a literature-based methodology, this phase leverages real-world data to identify patterns and explore housing market behavior using data analytics techniques. By shifting from a macro-level theoretical framework to a focused, applied analysis, this project aims to translate conceptual insights into actionable understanding bridging the gap between global sustainability goals and localized urban realities.

2.1.2 Housing, Urbanization, and the Sustainable Cities Agenda

In recent years, rapid urbanization and rising housing prices have significantly influenced how individuals think about where and how they live. Especially in highly competitive metropolitan areas, issues such as affordability, housing accessibility, and socio-spatial inequality have become increasingly urgent. These challenges are not only economic concerns but also deeply linked to the broader sustainability agenda.

Within the framework of the United Nations Sustainable Development Goals (SDGs), Goal 11: Sustainable Cities and Communities explicitly calls for inclusive, safe, resilient,

and sustainable urban environments. Housing lies at the very core of this vision. Affordable and adequate housing contributes to social equity, supports public health outcomes (SDG 3), and connects to infrastructure planning and innovation (SDG 9). In other words, housing functions as a convergence point where multiple SDGs intersect.

This project, therefore, positions housing as a strategic entry point for urban sustainability research. By applying data analytics techniques, the study aims to uncover patterns in pricing behavior, market heat, and regional disparities, providing insights that can inform urban planning and policymaking aligned with sustainability goals. The focus on housing allows for a localized, evidence-based exploration of sustainability challenges that are both measurable and socially impactful.

To guide this investigation, the following research questions were formulated:

- What are the key drivers of variation in housing prices across U.S. regions?
- How do affordability and market competitiveness indicators influence home sale prices?
- To what extent can machine learning models improve the prediction of housing prices compared to linear methods?
- What temporal trends can be observed in housing prices, and how might they evolve in the near future?

These questions provide the analytical foundation for the study and highlight the relevance of housing markets as a domain for sustainability-focused research.

2.2 Data Preprocessing and Cleaning

2.2.1 Data Sources and Collection Methods

The dataset used in this project was obtained from **Zillow Research**, a public data platform that provides comprehensive and regularly updated information on the U.S. housing market. It includes a wide range of variables such as median sale prices, rental prices, market heat indices, affordability metrics, and new construction data. These indicators are essential for identifying temporal patterns and regional disparities in housing trends.

Although the initial intention was to conduct a comparative housing analysis using data from Bucharest, access to this data was restricted due to real estate privacy limitations. As a result, Zillow's datasets were selected as the primary source because of their high coverage, reliability, and time-based consistency.

The analysis focuses on the dynamics of urban housing markets, particularly how variables such as affordability and market demand fluctuate over time. Given the ongoing housing affordability crisis in many U.S. cities, understanding these patterns has become increasingly important for researchers, urban planners, and policymakers alike.

To enable structured time-series analysis, seven individual datasets were merged into a unified panel. These include:

- Median Sale Price
- Percentage of Homes Sold Above List Price
- Median Rent (ZORI)
- Market Heat Index
- Affordability (Years to Save)
- Sale-to-List Price Ratio
- New Construction Median Sale Price

Each dataset was reshaped into long format and merged using common identifiers: `RegionID`, `RegionName`, `RegionType`, `StateName`, and `Date` with using Python. The final dataset comprises 10,500 rows and 12 columns, covering the period from **March 2018 to February 2025**.

Variable Descriptions:

- **RegionID:** Unique code for each geographic region.
- **RegionName:** Name of the city or metro area.
- **RegionType:** Type of region.
- **StateName:** U.S. state where the region is located.
- **Date:** Monthly timestamp of the data.
- **Median Sale Price:** Typical home sale price in the region.
- **Pct Sold Above List:** Percentage of homes sold above listing price, indicating market competitiveness.
- **Median Rent:** Typical rental price across listed homes.
- **Market Heat Index:** Zillow's indicator of how competitive the housing market is.
- **YearsToSave:** Estimated time needed for a household to save for a standard down payment.
- **SaleToListRatio:** Ratio of sale price to original list price, showing pricing dynamics.
- **NewCon Median Sale Price:** Median price of newly constructed homes in the region.

2.2.2 Quality control procedures

To ensure the reliability and consistency of the dataset, a series of quality control procedures were applied. These included handling missing values, detection of outliers, removal of duplicate records, and verification of data types. Each step was implemented to preserve the integrity of the analysis pipeline and maintain consistency across all variables. Missing values and duplicate entries were identified using Python.

The initial dataset contained 10,500 rows and 12 variables. After merging multiple datasets, each row represented housing data for a specific city and month. However, not all cities had data available for every variable. For example, not every region includes newly constructed homes, which explains the missing values in the `NewConMedianSalePrice` column.

Smaller amounts of missing data in variables such as `PctSoldAboveList`, `SaleToListRatio`, and `MedianRent` were handled by removing only the affected rows. Basic data type checks confirmed appropriate formatting in Python. A heatmap was used to visualize missing values across columns. The most significant portion of missing data appeared in `NewConMedianSalePrice` (1,718 rows, approximately 16%), which was retained in order to preserve valuable information for regions where new construction activity is reported.

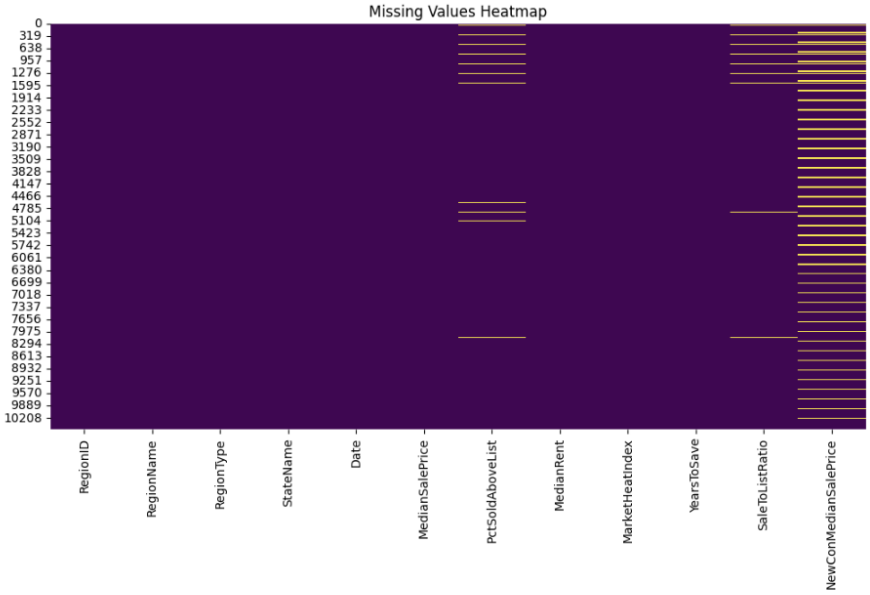
No duplicate records were detected. After the cleaning process, the dataset was reduced to 9,893 rows, ensuring consistency while maintaining as much valuable information as possible for further analysis.

```
Shape: (10500, 12)

 Column Types:
 RegionID                          int64
RegionName                        object
RegionType                        object
StateName                         object
Date                      datetime64[ns]
MedianSalePrice                  float64
PctSoldAboveList                 float64
MedianRent                       float64
MarketHeatIndex                  float64
YearsToSave                      float64
SaleToListRatio                  float64
NewConMedianSalePrice            float64
dtype: object

 Missing values per column:
 NewConMedianSalePrice    1718
PctSoldAboveList          301
SaleToListRatio           291
MedianRent                191
StateName                  84
dtype: int64
```

Figure 2.1: Data types and missing value summary for the merged dataset

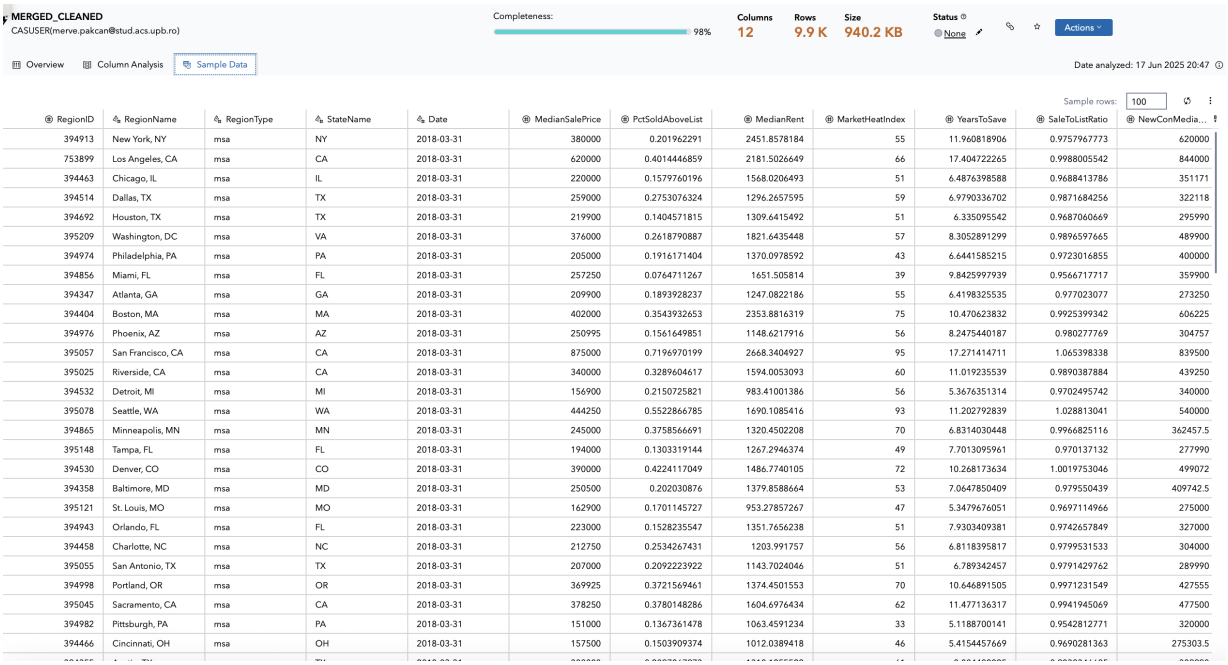



```
Rows before drop: 10500
Rows after drop (excluding only key variables): 9893

Duplicate Rows: 0
```

Figure 2.2: Missing values heatmap and post-cleaning row statistics

Following the merging and cleaning processes performed in Python, the dataset was imported into SAS via the Manage Data interface. Additionally, the Discover Information Assets part was utilized to explore summary statistics and gain a general overview of the dataset.The dependent variable used in this analysis is *MedianSalePrice*, representing housing market value. Independent variables include *MedianRent*, *NewConMedianSalePrice*, *PctSoldAboveList*, *SaleToListRatio*, *YearsToSave*, and *MarketHeatIndex*, as they are considered influential factors on home prices.



RegionID	RegionName	RegionType	StateName	Date	MedianSalePrice	PctSoldAboveList	MedianRent	MarketHeatIndex	YearsToSave	SaleToListRatio	NewConMedia…
394913	New York, NY	msa	NY	2018-03-31	380000	0.201962291	2451.8578184	55	11.960818906	0.9757967773	620000
753899	Los Angeles, CA	msa	CA	2018-03-31	620000	0.4014446859	2181.5026649	66	17.404722265	0.9988005542	844000
394463	Chicago, IL	msa	IL	2018-03-31	220000	0.1579760196	1568.0206493	51	6.4876398588	0.9688413786	351171
394514	Dallas, TX	msa	TX	2018-03-31	259000	0.2753076324	1296.2657595	59	6.9790336702	0.9871684256	322118
394692	Houston, TX	msa	TX	2018-03-31	219900	0.1404571815	1309.6415492	51	6.335095542	0.9687060669	295990
395209	Washington, DC	msa	VA	2018-03-31	376000	0.2618790887	1821.6435448	57	8.3052891299	0.9896597665	489900
394974	Philadelphia, PA	msa	PA	2018-03-31	205000	0.1916171404	1370.0978592	43	6.6441585215	0.9723016855	400000
394856	Miami, FL	msa	FL	2018-03-31	257250	0.0764711267	1651.505814	39	9.8425997939	0.9566717717	359900
394347	Atlanta, GA	msa	GA	2018-03-31	209900	0.1893928237	1247.0822186	55	6.4198325535	0.977023077	273250
394404	Boston, MA	msa	MA	2018-03-31	402000	0.3543932653	2353.8816319	75	10.470623832	0.9925399342	606225
394976	Phoenix, AZ	msa	AZ	2018-03-31	250995	0.1561649851	1148.6217916	56	8.2475440187	0.980277769	304757
395057	San Francisco, CA	msa	CA	2018-03-31	875000	0.7196970199	2668.3404927	95	17.271414711	1.065398338	839500
395025	Riverside, CA	msa	CA	2018-03-31	340000	0.3289604617	1594.0053093	60	11.019235539	0.9890387884	439250
394532	Detroit, MI	msa	MI	2018-03-31	156900	0.2150725821	983.41001386	56	5.3676351314	0.9702495742	340000
395078	Seattle, WA	msa	WA	2018-03-31	444250	0.5522866785	1690.1085416	93	11.202792839	1.028813041	540000
394865	Minneapolis, MN	msa	MN	2018-03-31	245000	0.3758566691	1320.4502208	70	6.8314030448	0.9966825116	362457.5
395148	Tampa, FL	msa	FL	2018-03-31	194000	0.1303319144	1267.2946374	49	7.7013095961	0.970137132	277990
394530	Denver, CO	msa	CO	2018-03-31	390000	0.4224117049	1486.7740105	72	10.268173634	1.0019753046	499072
394358	Baltimore, MD	msa	MD	2018-03-31	250500	0.202030876	1379.8588664	53	7.0647850409	0.979550439	409742.5
395121	St. Louis, MO	msa	MO	2018-03-31	162900	0.1701145727	953.27857267	47	5.3479676051	0.9697114966	275000
394943	Orlando, FL	msa	FL	2018-03-31	223000	0.1528235547	1351.7656238	51	7.9303409381	0.9742657849	327000
394458	Charlotte, NC	msa	NC	2018-03-31	212750	0.2534267431	1203.991757	56	6.8118395817	0.9799531533	304000
395055	San Antonio, TX	msa	TX	2018-03-31	207000	0.2092223922	1143.7024046	51	6.789342457	0.9791429762	289990
394998	Portland, OR	msa	OR	2018-03-31	369925	0.3721569461	1374.4501553	70	10.646891505	0.9971231549	427555
395045	Sacramento, CA	msa	CA	2018-03-31	378250	0.3780148286	1604.6976434	62	11.477136317	0.9941945069	477500
394982	Pittsburgh, PA	msa	PA	2018-03-31	151000	0.1367361478	1063.4591234	33	5.1188700141	0.9542812771	320000
394466	Cincinnati, OH	msa	OH	2018-03-31	157500	0.1503909374	1012.0389418	46	5.4154457669	0.9690281363	275303.5

Figure 2.4: Sas Discover Information Assets

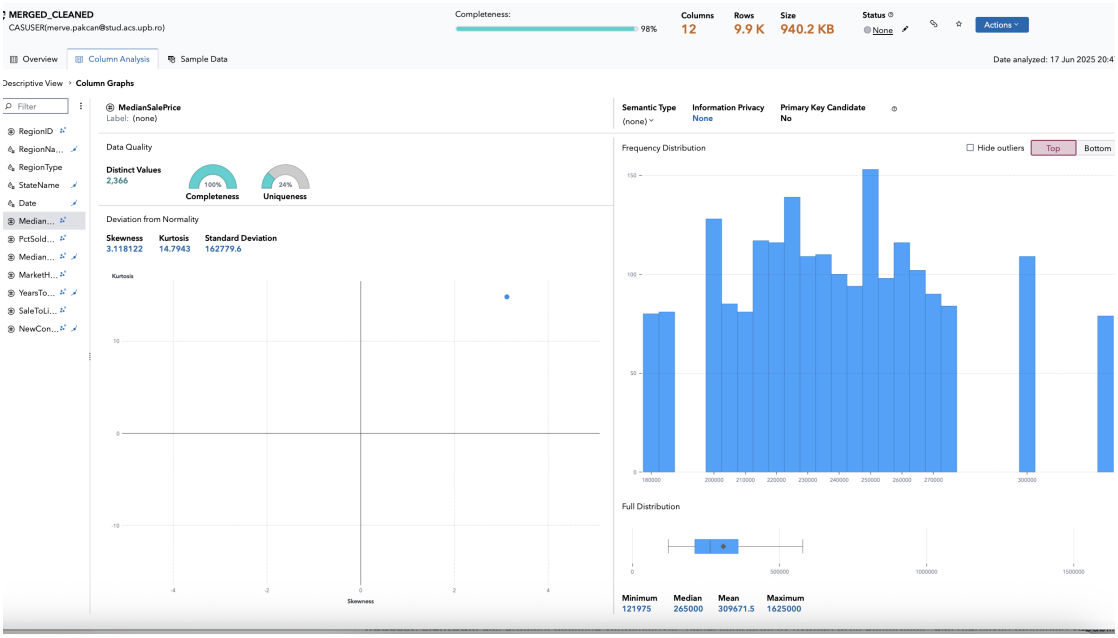
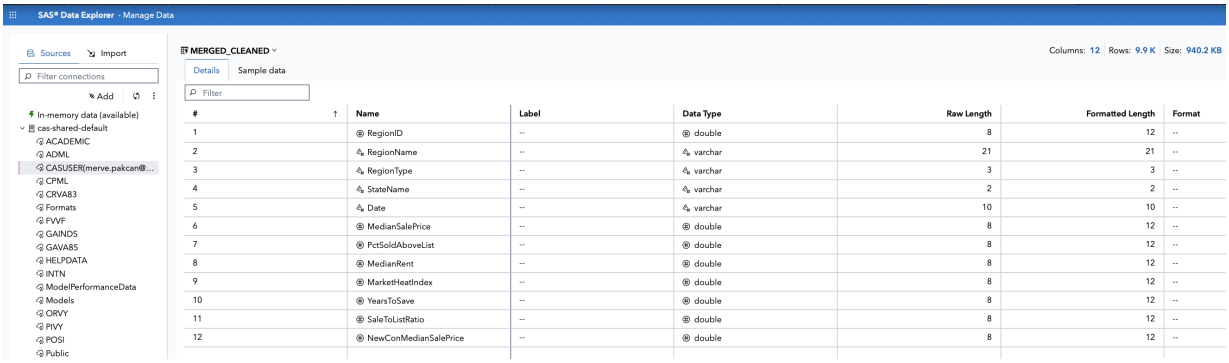


Figure 2.3: Column analysis of the cleaned and merged Zillow dataset

The dataset comprises 9,900 rows and 12 columns, with a total size of 940.2 KB. Out of these 12 columns, 8 are numerical (DataType: double) and 4 are character data (DataType: varchar), offering a robust data structure for analysis.



#	Name	Label	Data Type	Raw Length	Formatted Length	Format
1	RegionID	--	double	8	12	--
2	RegionName	--	varchar	21	21	--
3	RegionType	--	varchar	3	3	--
4	StateName	--	varchar	2	2	--
5	Date	--	varchar	10	10	--
6	MedianSalePrice	--	double	8	12	--
7	PctSoldAboveList	--	double	8	12	--
8	MedianRent	--	double	8	12	--
9	MarketHeatIndex	--	double	8	12	--
10	YearsToSave	--	double	8	12	--
11	SaleToListRatio	--	double	8	12	--
12	NewConMedianSalePrice	--	double	8	12	--

Figure 2.5: Data type summary of the cleaned dataset in SAS Data Explorer

Also, after loading the data into SAS environment, data types controlled in Manage Data part.

Outlier detection was performed using the Explore and Visualize tools in SAS. Box plots were generated for key numerical variables to examine the distribution and identify any extreme values. Rather than displaying individual data points, SAS visualizations use gradient-shaded areas to indicate regions where outliers are concentrated. This approach provides a clear understanding of variability while maintaining a clean and interpretable visual layout.

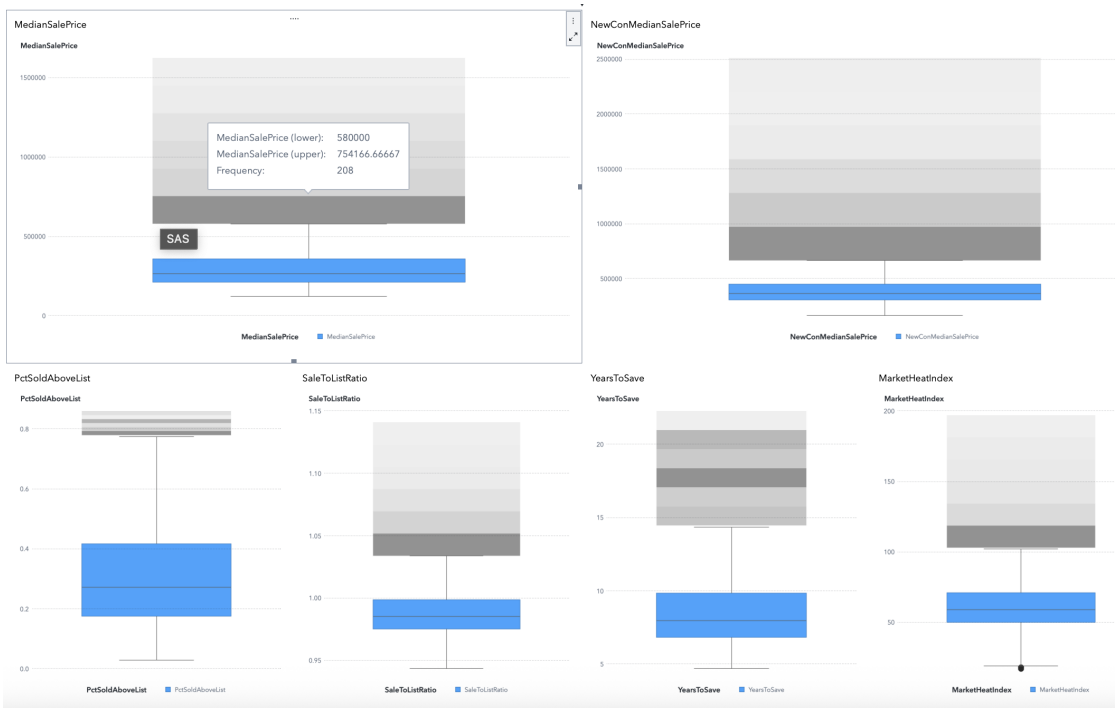


Figure 2.6: Box plot for `MedianSalePrice`

The variables *MedianSalePrice* and *NewConMedianSalePrice* exhibited right-skewed distributions with clearly defined upper outlier zones. These outliers were particularly prominent in high-cost urban regions, reflecting actual market conditions rather than anomalies.

PctSoldAboveList and *SaleToListRatio* also showed high-value outliers, consistent with competitive market dynamics where homes frequently sell above their list prices. These patterns are particularly visible through the shaded bands above the box range.

In the case of *YearsToSave* and *MarketHeatIndex*, outliers indicated regions where affordability and buyer competition significantly diverged from the national median. These values were not removed, as they represent meaningful variations in housing market behavior.

Retaining such outliers was considered essential, as they capture important aspects of real-world market diversity and contribute to the explanatory power of the dataset in subsequent analyses.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) serves as a foundational stage in data-driven research, aiming to uncover patterns, detect anomalies, and guide subsequent modeling through graphical and statistical techniques. As emphasized by Tukey (1977), EDA is not merely

a preliminary step but a critical process for developing meaningful insights from complex datasets Tukey et al., 1977.

2.3.1 Descriptive Statistics

	RegionID	Date	MedianSalePrice	\
count	9893.000000	9893	9.893000e+03	
mean	406601.992621	2021-10-16 19:09:28.038006528	3.096715e+05	
min	394312.000000	2018-03-31 00:00:00	1.219750e+05	
25%	394531.000000	2020-01-31 00:00:00	2.131770e+05	
50%	394792.000000	2021-10-31 00:00:00	2.650000e+05	
75%	395005.000000	2023-07-31 00:00:00	3.599000e+05	
max	753912.000000	2025-02-28 00:00:00	1.625000e+06	
std	64112.991749	NaN	1.627796e+05	

	PctSoldAboveList	MedianRent	MarketHeatIndex	YearsToSave	\
count	9893.000000	9893.000000	9893.000000	9893.000000	
mean	0.306217	1500.371539	61.627211	8.631004	
min	0.029102	806.913837	17.000000	4.670567	
25%	0.175565	1208.192870	50.000000	6.822396	
50%	0.271981	1402.893622	59.000000	7.959976	
75%	0.416761	1715.942181	71.000000	9.836859	
max	0.859512	3347.407630	197.000000	22.269278	
std	0.163279	429.354211	17.170322	2.625767	

	SaleToListRatio	NewConMedianSalePrice
count	9893.000000	8.481000e+03
mean	0.988815	4.059965e+05
min	0.943317	1.620000e+05
25%	0.975242	3.062750e+05
50%	0.985311	3.650000e+05
75%	0.998708	4.499900e+05
max	1.140964	2.510000e+06
std	0.020732	1.759764e+05

Figure 2.7: Summary statistics of the key variables in the Zillow dataset

Table 2.7 provides a statistical overview of the main variables used in the analysis. The `MedianSalePrice` shows substantial variation across regions, with a mean of approximately $309,000 and a maximum value exceeding $1.6 million. New construction properties (`NewConMedianSalePrice`) reach up to $2.5 million in some cities, indicating high-end market segments.

The `PctSoldAboveList` variable, with an average of 36%, suggests that a significant portion of homes are sold above the listing price—an indication of competitive housing markets in certain regions.

Rental prices (`MedianRent`) range from $806 to over $3,300, while the `MarketHeatIndex` spans from 17 to 197, reinforcing the presence of stark regional differences in demand and buyer activity.

Notably, the `YearsToSave` variable, which estimates the number of years required to save for a down payment, averages at 8.6 years and reaches beyond 22 years in some areas. This underscores significant affordability challenges across cities.

These preliminary statistics offer key insights into housing price dynamics, regional competitiveness, and financial accessibility, setting a solid foundation for further exploratory analysis.

2.3.2 Distribution Analysis via Histograms

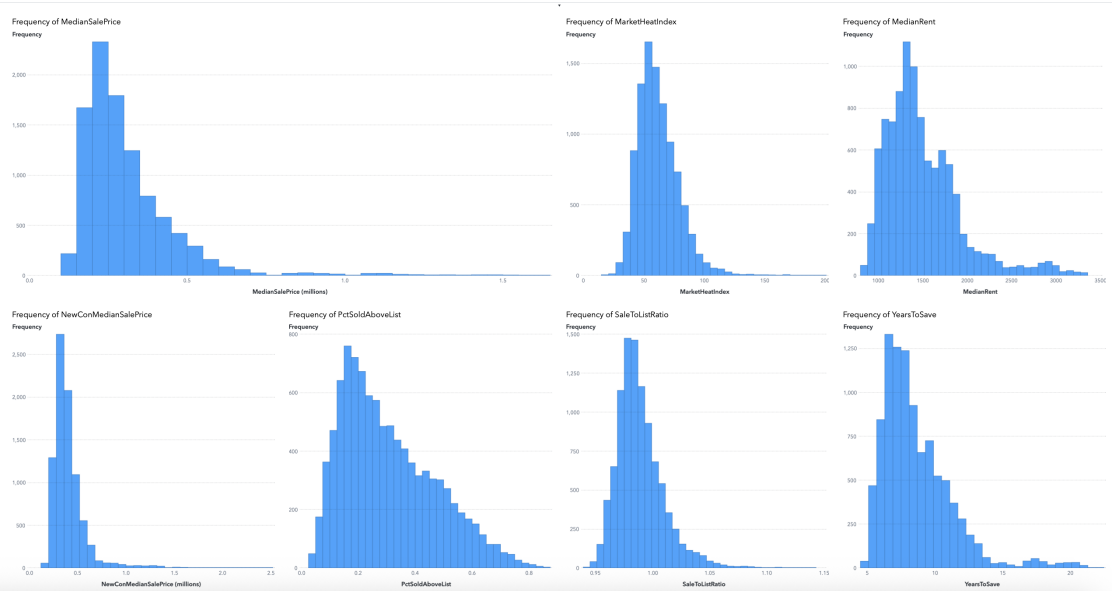


Figure 2.8: Distribution of key Zillow housing indicators.

The distribution of key numerical variables was examined using histograms generated in the SAS Visual Analytics environment. These plots provided a clear visual summary of the frequency and skewness patterns across the dataset.

Variables such as *MedianSalePrice*, *NewConMedianSalePrice*, and *MedianRent* demonstrated strong right-skewed distributions. This is consistent with real-world housing markets, where a majority of properties cluster around mid-range values, but a smaller portion of high-value listings extend the upper tail. These skewed patterns highlight the price diversity and economic inequality present in different housing regions.

Conversely, *MarketHeatIndex* followed a near-normal distribution, indicating that most markets experience moderate buyer competition, with fewer cases showing exceptionally high or low heat levels. *PctSoldAboveList* and *SaleToListRatio* also revealed slight right skewness, supporting the presence of aggressive market conditions where homes often sell above their listing price.

YearsToSave showed a right-skewed profile as well, with the majority of regions requiring a moderate savings period, while a smaller number of areas typically high-cost urban centers—demand significantly longer timeframes to afford a median-priced home.

Overall, the histograms validate the presence of non-normality and economic variability within the dataset. These characteristics provide valuable context for model development and interpretation in subsequent stages of the analysis.

2.3.3 Correlation Analysis of Housing Indicators

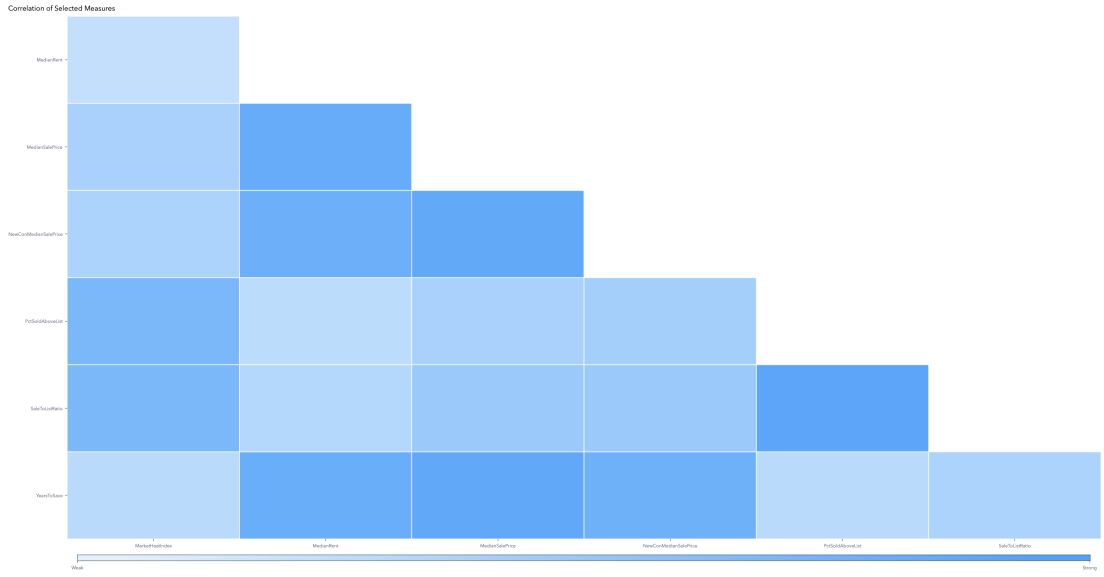


Figure 2.9: Correlation matrix

The correlation matrix illustrates the strength and direction of linear relationships among selected housing variables. As expected, strong positive correlations are observed between *MedianSalePrice*, *MedianRent*, and *YearsToSave*, suggesting that higher home prices are generally associated with higher rental costs and longer saving periods. Similarly, the strong association between *PctSoldAboveList* and *SaleToListRatio* reflects market competitiveness in overheated regions. In contrast, variables such as *MarketHeatIndex* show more moderate correlations, indicating their distinct but complementary role in explaining housing dynamics.

2.3.4 Q-Q Plots for Normality Assessment

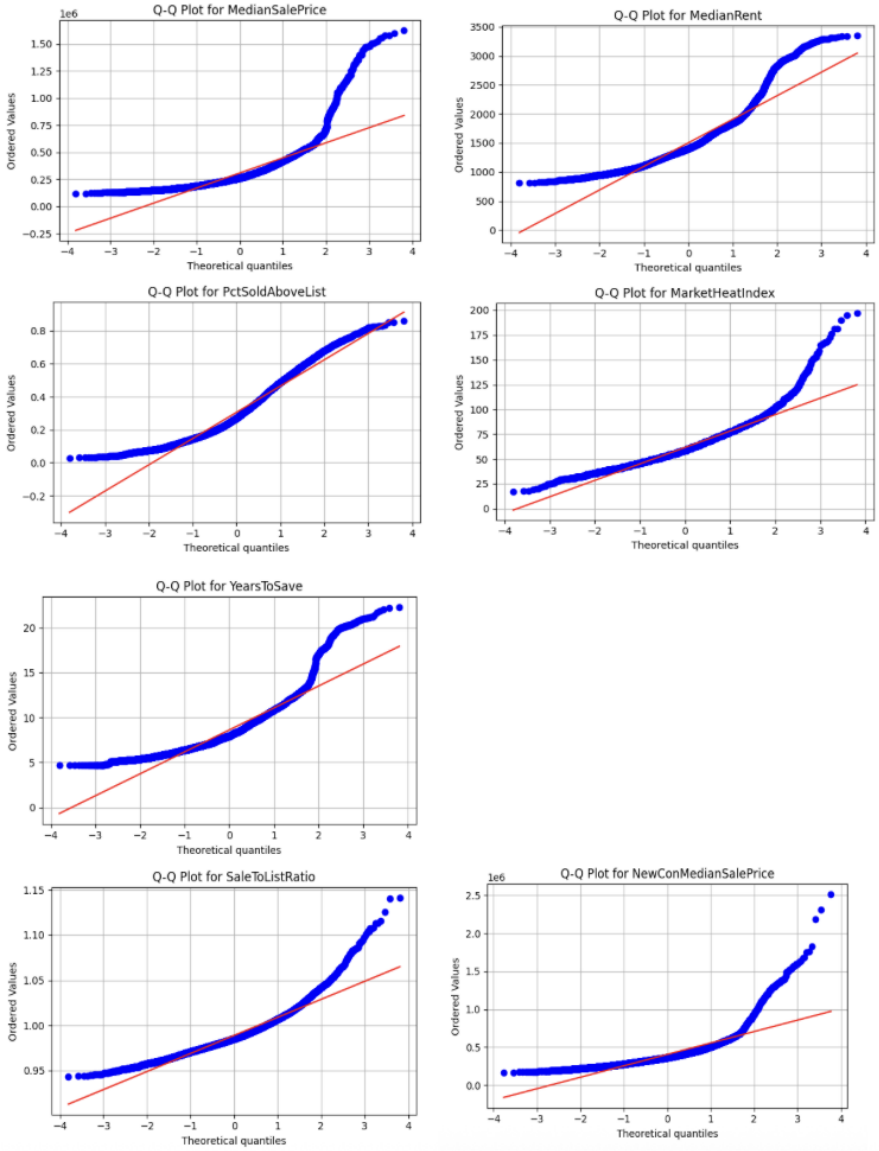


Figure 2.10: Q-Q plots

Q–Q plots were used to assess the normality of key numerical variables in the housing dataset. Most variables, including `MedianSalePrice`, `MedianRent`, and `YearsToSave`, display clear deviations from the theoretical reference line, particularly at the tails. This indicates strong right-skewness and the presence of extreme values.

For example, `MedianSalePrice` and `MedianRent` exhibit pronounced curvature at

both ends of the distribution, reflecting a small number of regions with exceptionally high prices. Similarly, variables such as `PctSoldAboveList` and `MarketHeatIndex` show S-shaped or curved patterns, suggesting departures from normality and potential clustering.

These distributional patterns confirm that the normality assumption does not hold for most features. Therefore, non-parametric or distribution-agnostic methods are adopted in the following analysis to ensure statistical robustness and valid inference.

2.3.5 Time Series Analysis of Rent and Sale Prices

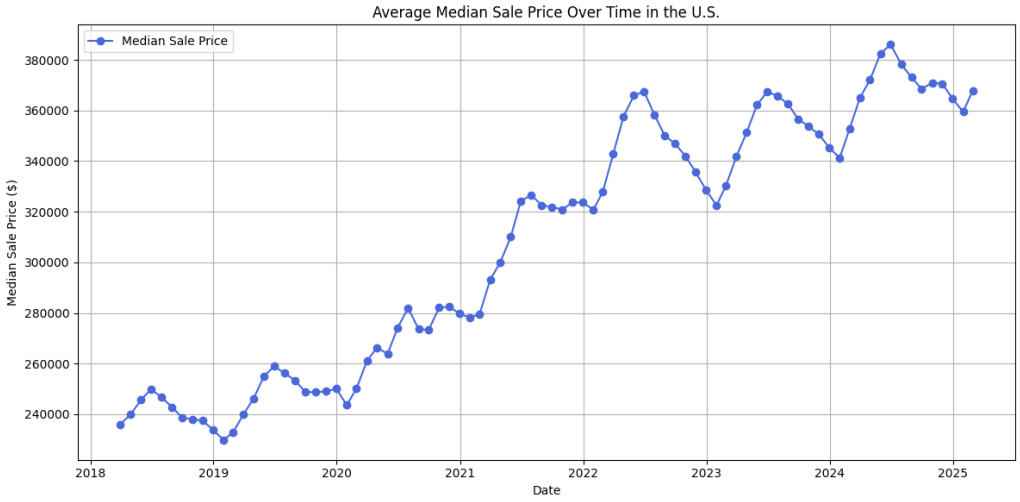


Figure 2.11: Time series plots of average monthly Median Sale Price

To capture long-term trends in housing dynamics, time series plots were generated for average monthly values of median sale price and median rent in the U.S. between 2018 and early 2025. Median Sale Price shows a steady upward trend with minor fluctuations. A significant acceleration is observed around 2021–2022, potentially reflecting post-pandemic housing demand surges and low interest rates during that period.

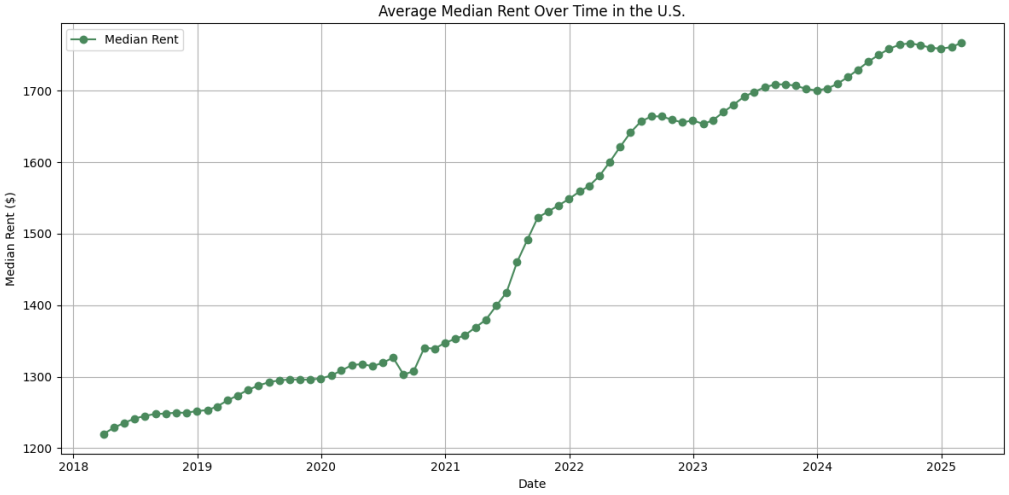


Figure 2.12: Time series plots of average monthly Median Rent Price

Median Rent follows a similar pattern, increasing gradually before rising more sharply from late 2021 onward. This may indicate increasing rental pressure in parallel with rising home prices. These temporal patterns suggest a strong cyclical growth in housing costs over recent years, reinforcing concerns around affordability and supporting the need for sustainable urban planning approaches.

2.3.6 Regional Analysis

To better illustrate the substantial differences between median sale prices and rents across states, the data was sorted in descending order. This highlights the regional disparities more clearly.

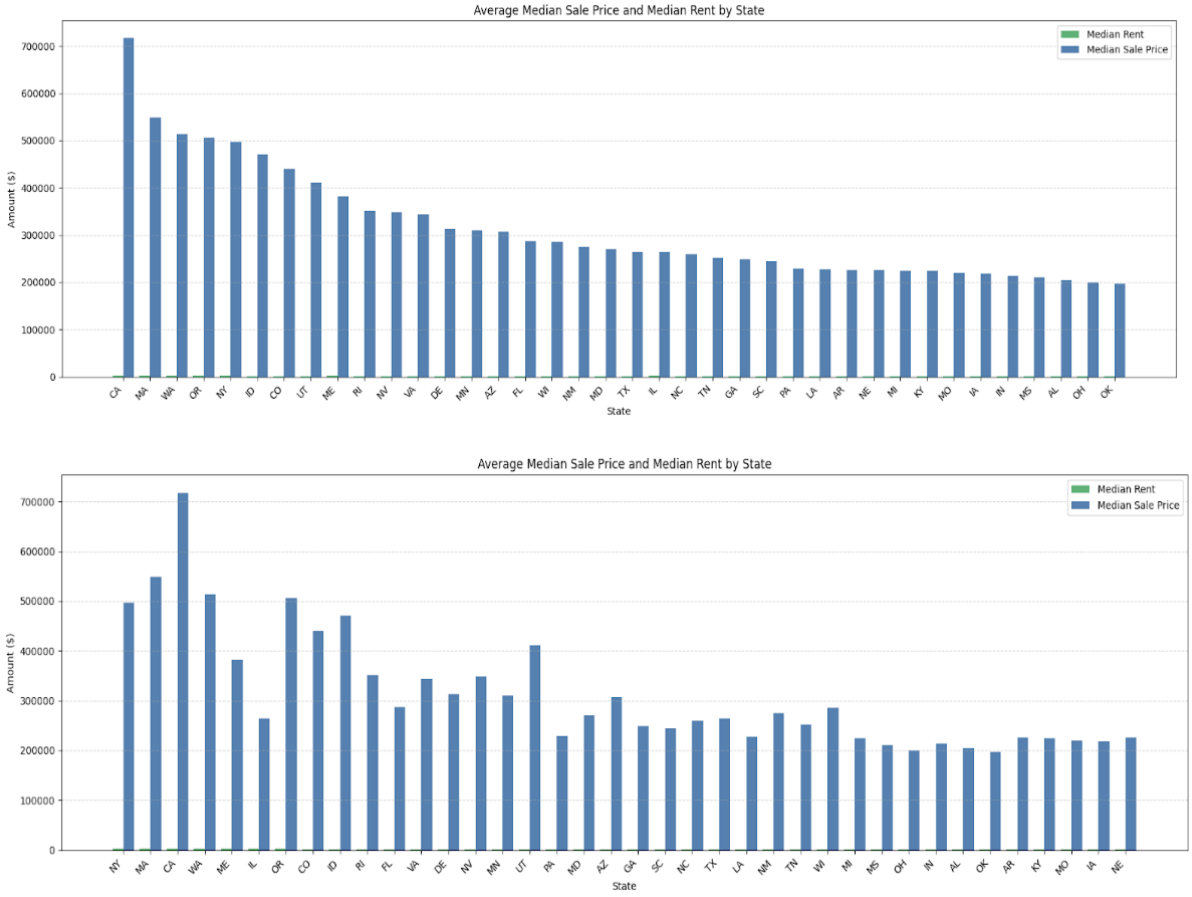


Figure 2.13: Comparison of average median sale price and median rent across U.S. states.

California has the highest median home sale price, followed by Washington and Massachusetts, all of which also report relatively high rental costs. Notably, New York has the highest median rent but does not appear among the top three states for sale prices.

In states like Idaho and Utah, home prices are comparatively high while rents remain moderate, resulting in a significant gap between ownership and rental markets. In contrast, Southern states such as Oklahoma and Mississippi exhibit both low sale prices and low rents, reflecting more affordable housing conditions.

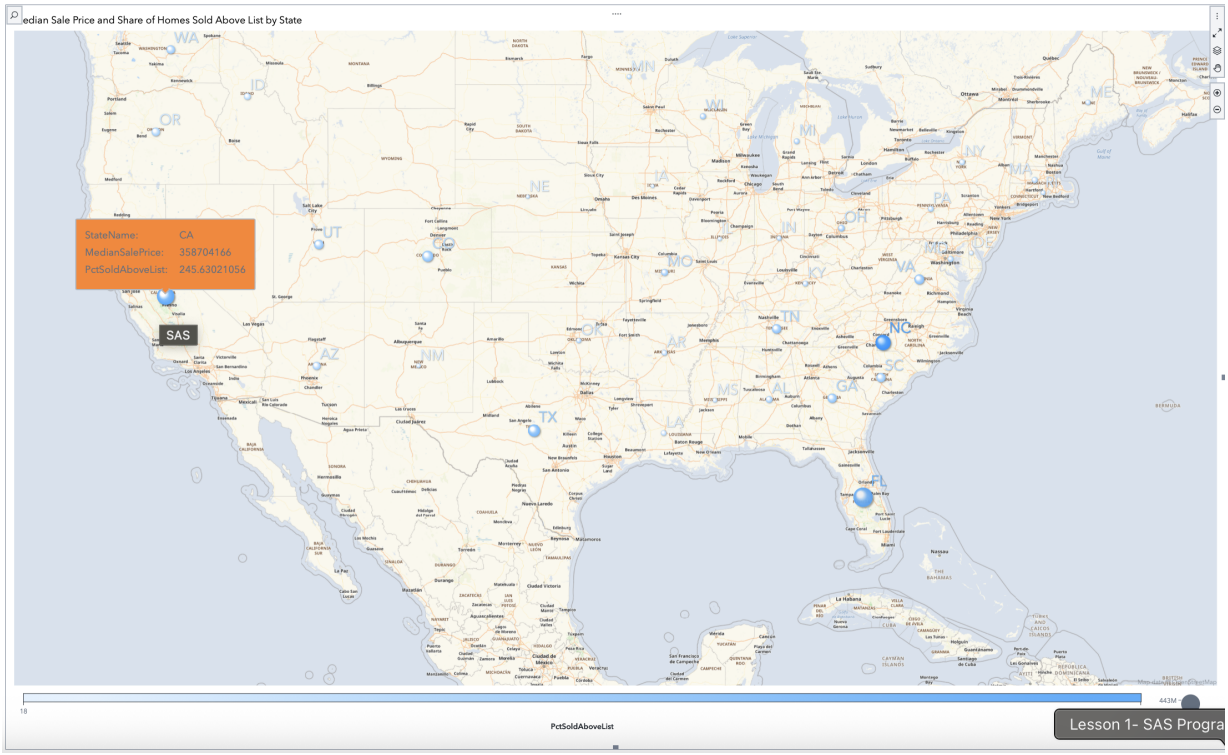


Figure 2.14: Geo Bubble Map showing Median Sale Price (size) and `PctSoldAboveList` (color) across U.S. states

Geographic Distribution of Median Sale Price and Market Competitiveness

The Geo Bubble Map provides a spatial overview of the U.S. housing market by visualizing the Median Sale Price (bubble size) and the PctSoldAboveList ratio (color intensity) across states. Larger bubbles indicate higher average sale prices, while more saturated blue tones reflect greater market competitiveness, a higher share of homes sold above the listing price. States like California (CA) and Florida (FL) exhibit both high prices and intense buyer competition, highlighting highly active housing markets. This visualization supports regional comparison and reveals that price and competitiveness often concentrate in coastal and metropolitan areas.

2.4 Feature Engineering and Model Design

To enhance the analytical depth of this study, three derived features were engineered from the original Zillow housing dataset using SAS Visual Analytics:

- **RentToPriceRatio:** Calculated as `MedianRent` divided by `MedianSalePrice`, this metric serves as a proxy for affordability and rental yield. It reflects the relative cost of renting versus owning a home.

- **MarketTension:** Constructed by multiplying `MarketHeatIndex` with `PctSoldAboveList`, this composite indicator captures latent demand and market competitiveness, highlighting regions with stronger upward pricing pressures.
- **LogMedianSalePrice:** This is the logarithm of `MedianSalePrice`. It is used to reduce skewness and stabilize variance, thereby improving model interpretability and performance in regression-based analyses.

These engineered features aim to better capture hidden market dynamics and are later used in exploratory analyses and predictive modeling.

To examine the factors influencing housing prices, a series of predictive models were constructed using SAS Visual Analytics. The aim was to evaluate how key indicators affect home sale prices across different U.S. regions.

The modeling process began with linear regression to establish a baseline and uncover interpretable relationships. This was followed by tree-based methods, including decision trees and ensemble models, to better capture non-linear interactions within the data.

Applying multiple modeling techniques enabled a comprehensive comparison of predictive accuracy, assessment of variable importance, and validation of results across approaches. Throughout the analysis, `LogMedianSalePrice` was used as the target variable to reduce skewness and improve both model fit and interpretability.

2.4.1 Linear Regression

Linear regression was used as a baseline model to understand the direct relationships between predictors and the target variable.

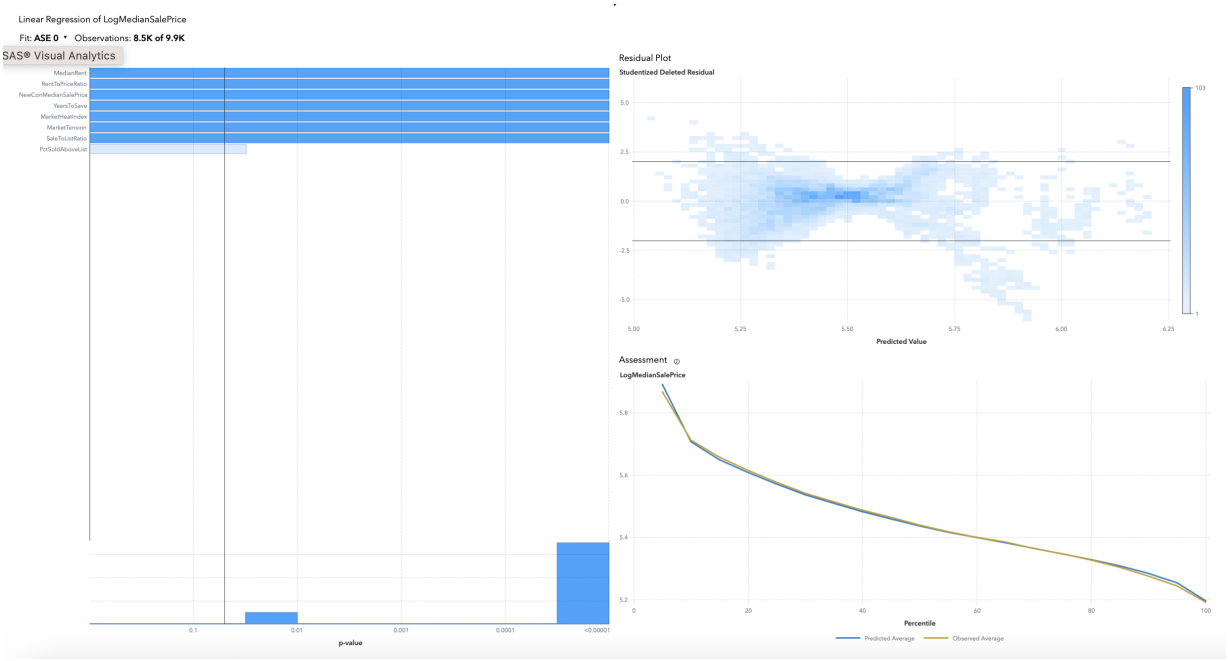


Figure 2.15: Linear regression model

The linear regression model indicated statistically significant associations between the transformed target variable, key predictors such as `MedianRent`, `RentToPriceRatio`, `NewConMedianSalePrice`, and `YearsToSave`. The assessment curve showed strong alignment between predicted and observed values across percentiles, supporting the model's validity as a baseline approach.

However, the residual plot suggested non-linearity and signs of heteroscedasticity, with residuals spreading unevenly across predicted values. These limitations highlighted the linear model's inability to fully capture market complexity. As a result, more adaptive, non-linear models were pursued in subsequent analysis stages.

2.4.2 Random Forest

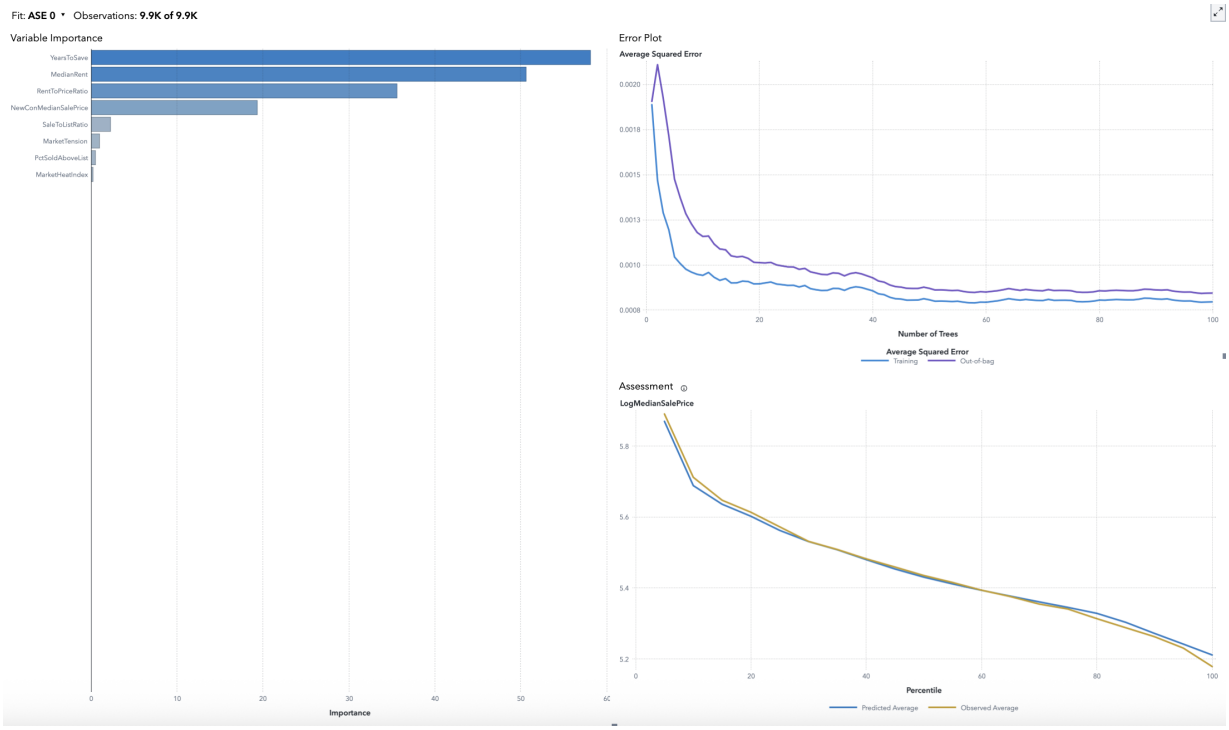


Figure 2.16: Random forest model

The random forest model demonstrated strong predictive accuracy for `LogMedianSalePrice`, with low out-of-bag error and strong agreement between predicted and observed values across all percentiles. The error plot confirmed that training and validation errors converged steadily, suggesting a stable and generalizable model with low risk of overfitting.

`YearsToSave`, `MedianRent`, and `RentToPriceRatio` emerged as the top contributing features in the variable importance chart. Compared to the linear model, the random forest provided enhanced flexibility and captured more complex relationships, improving overall predictive power and robustness.

2.4.3 Decision Tree

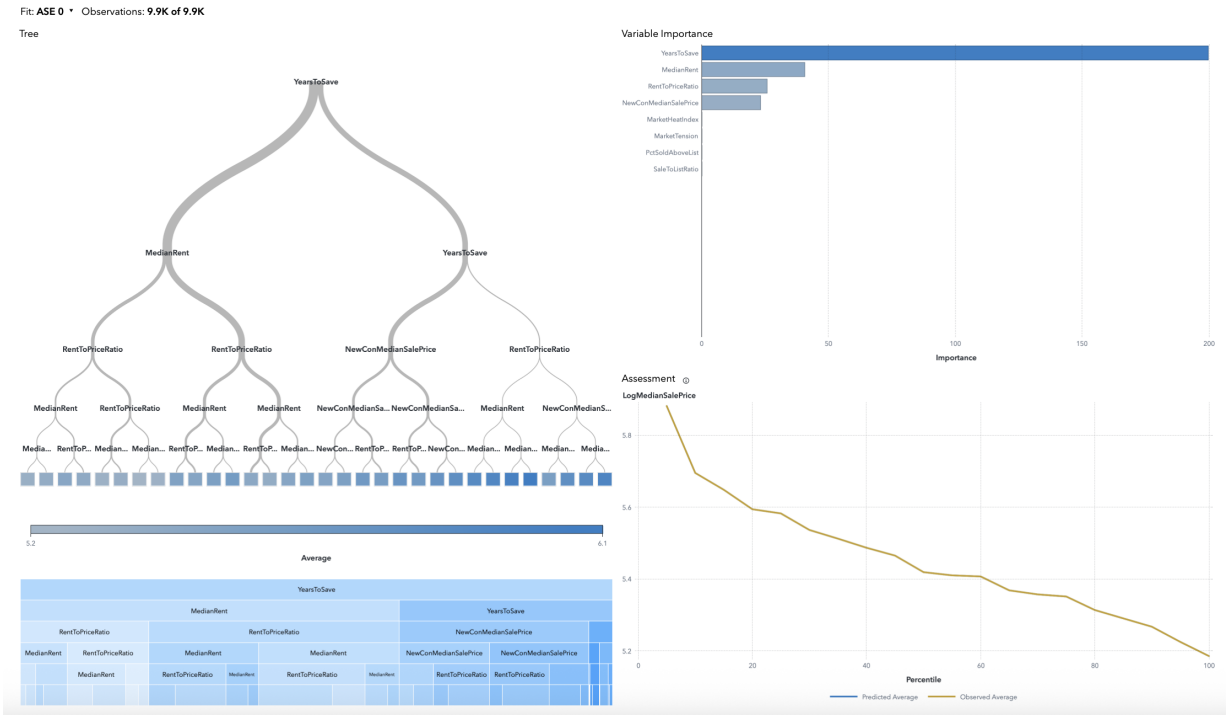


Figure 2.17: Decision tree model

The decision tree model showed that `YearsToSave` was the most important variable for predicting `LogMedianSalePrice`, followed by `MedianRent` and `RentToPriceRatio`. The tree made clear splits based on affordability, making the structure easy to follow and interpret.

Even though the model picked up on meaningful patterns, the assessment curve pointed to some underfitting, likely due to the limits of using a single tree. Still, it helped reveal how the variables relate to each other and offered a good view of their overall influence.

2.4.4 Gradient Boosting

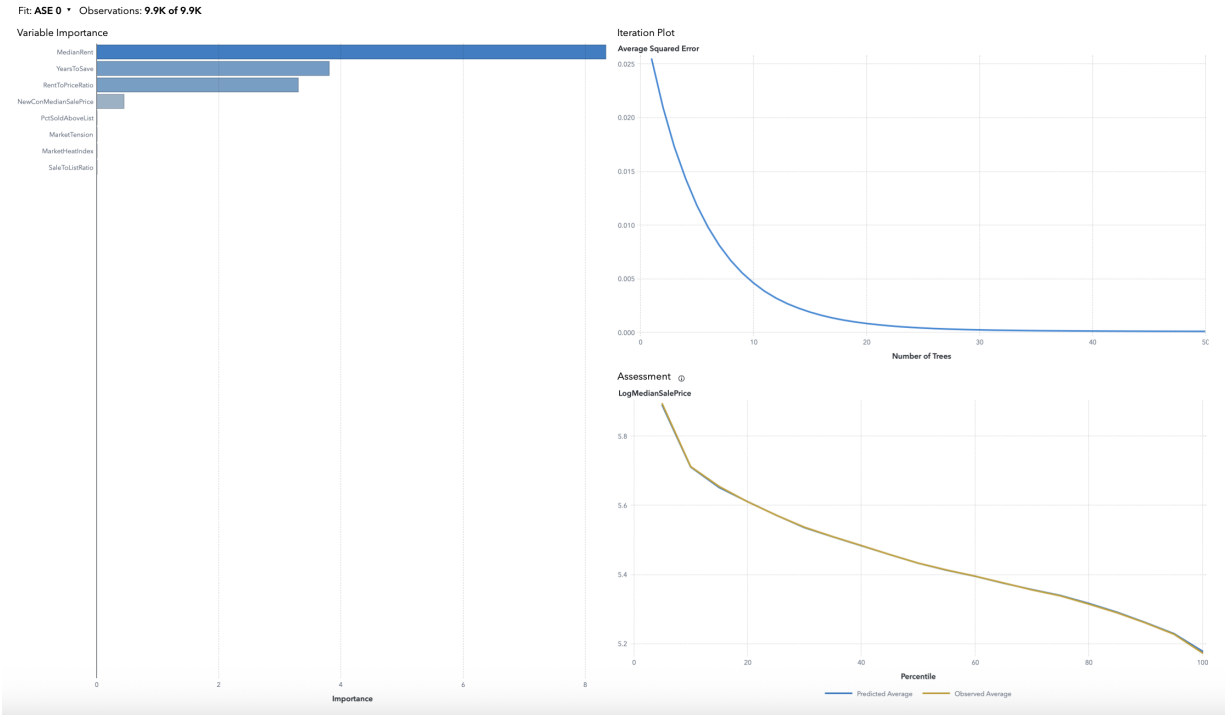


Figure 2.18: Gradient boosting model

The gradient boosting model showed strong accuracy. The error dropped quickly and stabilized early. The most influential predictors were `MedianRent`, `YearsToSave`, and `RentToPriceRatio`. Other variables had minimal impact, which could be due to feature overlap or reduced non-linear effects after log transformation. The predicted values aligned closely with the observed values across percentiles.

2.4.5 Model Comparison

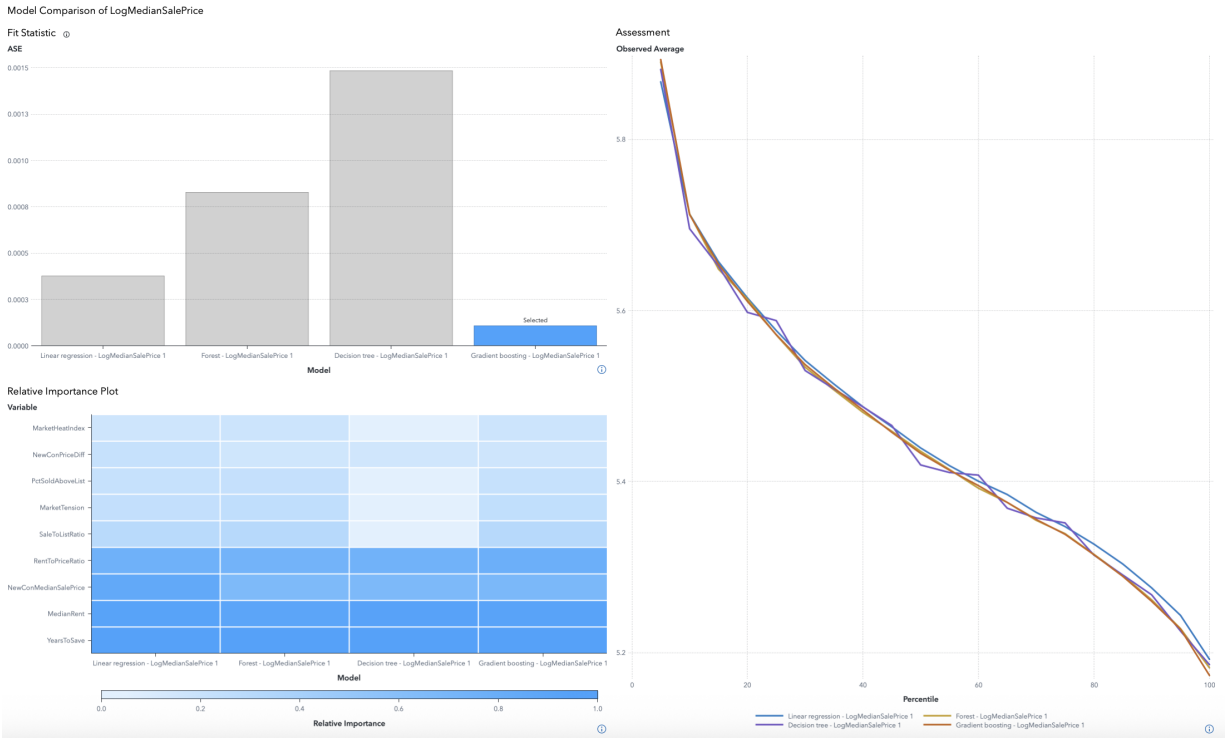


Figure 2.19: Model comparison

Among the evaluated models, **Gradient Boosting** achieved the lowest ASE, confirming its superior predictive accuracy. **Random Forest** followed closely, while **Linear Regression** provided a solid baseline. The **Decision Tree**, although interpretable, showed the weakest performance.

Assessment curves showed that Gradient Boosting most closely aligned with the observed values, especially across middle and upper percentiles, reflecting robust generalization.

Across all models, `YearsToSave` and `MedianRent` consistently ranked among the top predictors. This reinforces their strong influence on `LogMedianSalePrice` and highlights their importance in housing market dynamics.

2.4.6 Future Price Projection Using Time Series Analysis

To add a future-looking element to the project, a time series forecast was made using the Prophet library in Python. The goal was to take a first step toward including time-based trends in housing price analysis. Using data from 2018 to 2025, the model predicted median sale prices for the next two years.

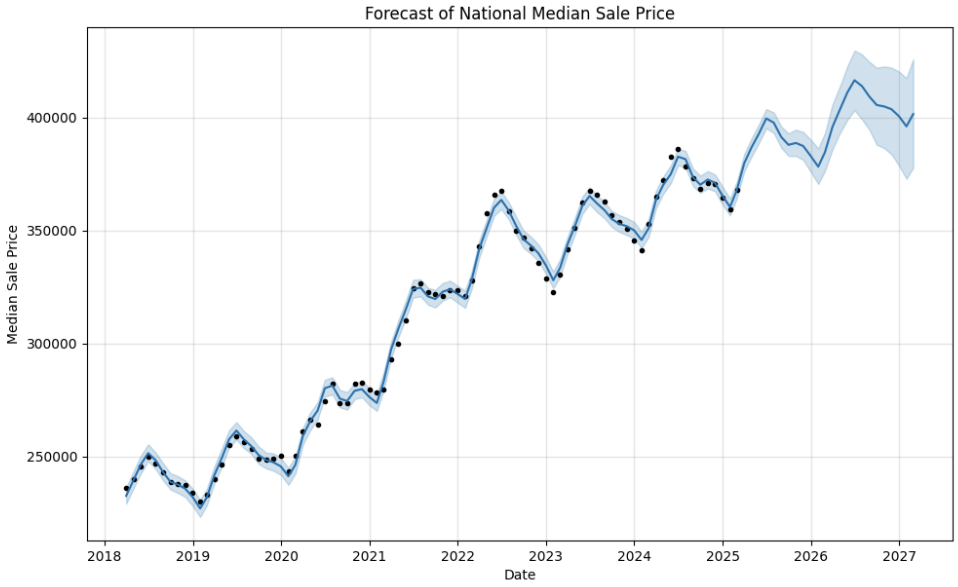


Figure 2.20: Prophet

As it seen from the graph, the forecast shows a steady upward trend, though the confidence interval gets wider over time. This suggests more uncertainty further out. Still, the general direction fits with the pattern that is seen in the past data. While this was just a basic attempt, it shows that time series models can be useful alongside other methods to better understand housing market trends.A detailed analysis will be carried out in the next semester.

2.5 Conclusion

This project examined the main factors influencing housing prices in U.S. cities by combining data preparation, exploratory analysis, and predictive modeling. Using Zillow's dataset covering the years 2018 to 2025, the project focused on how variables related to affordability, rental pressure, and market activity shape price levels across regions.

Exploratory data analysis helped uncover initial patterns through distribution plots, summary statistics, and correlation heatmaps. These steps highlighted key relationships and regional differences, which guided the selection of features for modeling.

Linear regression was used as a starting point, followed by decision trees, random

forests, and gradient boosting to capture more complex patterns. Among all, gradient boosting showed the best predictive performance. Across models, `YearsToSave` and `MedianRent` consistently stood out as the most important features.

To support a forward-looking perspective, a simple forecast using the Prophet model in Python was added. This time series projection extended the analysis by showing expected future price trends based on recent historical data.

Overall, the project shows how different data science techniques can be used together to understand urban housing dynamics. The results emphasize the role of affordability and competition in shaping housing prices and provide a strong basis for future work in real estate analysis and urban policy.

Appendices

Appendices

Appendix A

Fișiere sursă

Listing A.1: Cod Matlab – fișier complet

```
function [varargout]=drawCells(H,h,tuples,varargin)

[~,d]=size(H);

if nargin<4
    opt={'Alpha',0.4,'Color','b'};
else
    opt={varargin{:}{:}};
end

tmp_all=[];
for i=1:size(tuples,2)
    tuple=tuples(:,i);
    ii=find(tuple~=0.5); % discard the indices corresponding to hyperplanes wi
    tmp=Polyhedron(H(ii,:).*repmat(2*tuple(ii)-1,1,d),h(ii).*(2*tuple(ii)-1));
    tmp_all=[tmp_all tmp];
end

switch nargout
    case 1
        varargout{1}=tmp_all;
    case 0
        a=axis(gca);
        hold on
        plot(tmp_all,opt{:})
        axis(a);
    case 2
        varargout{1}=tmp_all;
```

```
        a=axis(gca);
        hold on
        varargout{2}=plot(tmp_all,opt{:});
        axis(a);
    otherwise
        error 'not an accepted number of outputs'
end
```

Listing A.2: Cod Matlab – fragment de fișier

```
    case 0
        error 'you need at least one argument'
    case 1
        H=varargin{1};
        [N,d]=size(H);
        if d>3
            error 'space dimension is too large'
        end
        h=ones(N,1);
        a=repmat([-1 1],1,d);
    case 2
```

Bibliography

Barbier, E. B., & Burgess, J. C. (2019). Sustainable development goal indicators: Analyzing trade-offs and complementarities. *World development*, *122*, 295–305.

Bennich, T., Weitz, N., & Carlsen, H. (2020). Deciphering the scientific literature on sdg interactions: A review and reading guide. *Science of the Total Environment*, *728*, 138405.

Dang, H.-A. H., & Serajuddin, U. (2020). Tracking the sustainable development goals: Emerging measurement challenges and further reflections. *World Development*, *127*, 104570.

Fuso Nerini, F., Tomei, J., To, L. S., Bisaga, I., Parikh, P., Black, M., Borrion, A., Spataru, C., Castán Broto, V., Anandarajah, G., et al. (2018). Mapping synergies and trade-offs between energy and the sustainable development goals. *Nature Energy*, *3*(1), 10–15.

Griggs, D., Nilsson, M., Stevance, A., McCollum, D., et al. (2017). *A guide to sdg interactions: From science to implementation*. International Council for Science, Paris.

Hák, T., Janoušková, S., & Moldan, B. (2016). Sustainable development goals: A need for relevant indicators. *Ecological indicators*, *60*, 565–573.

Horvath, S.-M., Muhr, M. M., Kirchner, M., Toth, W., Germann, V., Hundscheid, L., Vacik, H., Scherz, M., Kreiner, H., Fehr, F., et al. (2022). Handling a complex agenda: A review and assessment of methods to analyse sdg entity interactions. *Environmental Science & Policy*, *131*, 160–176.

Luchian, A., Drăgoicea, M., Fux, A., & Rozenes, S. (2025). Harnessing data analysis for global sustainability: An integrated approach to environmental, economic, and social wellbeing. *West, S., Meierhofer, J. Smart Services Summit 2024. Progress in IS*.

Lusseau, D., & Mancini, F. (2019). Income-based variation in sustainable development goal interaction networks. *Nature Sustainability*, *2*(3), 242–247.

Meng, X.-L. (2021). What are the values of data, data science, or data scientists?

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*, t. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of internal medicine*, *151*(4), 264–269.

Moyer, J. D., & Bohl, D. K. (2019). Alternative pathways to human development: Assessing trade-offs and synergies in achieving the sustainable development goals. *Futures*, *105*, 199–210.

Nilashi, M., Keng Boon, O., Tan, G., Lin, B., & Abumalloh, R. (2023). Critical data challenges in measuring the performance of sustainable development goals: Solutions and the role of big-data analytics. *Harvard Data Science Review*, *5*(3), 1–36.

Nilsson, M., Chisholm, E., Griggs, D., Howden-Chapman, P., McCollum, D., Messerli, P., Neumann, B., Stevance, A.-S., Visbeck, M., & Stafford-Smith, M. (2018). Mapping interactions between the sustainable development goals: Lessons learned and ways forward. *Sustainability science*, *13*, 1489–1503.

Pradhan, P., Costa, L., Rybski, D., Lucht, W., & Kropp, J. P. (2017). A systematic study of sustainable development goal (sdg) interactions. *Earth's Future*, *5*(11), 1169–1179.

Salleh, K. A., & Janczewski, L. (2019). Security considerations in big data solutions adoption: Lessons from a case study on a banking institution. *Procedia Computer Science*, *164*, 168–176.

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., & Lockwood, C. (2021). How to properly use the PRISMA statement. *Systematic Reviews*, *10*, 1–3.

Scharlemann, J. P., Brock, R. C., Balfour, N., Brown, C., Burgess, N. D., Guth, M. K., Ingram, D. J., Lane, R., Martin, J. G., Wicander, S., et al. (2020). Towards understanding interactions between sustainable development goals: The role of environment–human linkages. *Sustainability science*, *15*, 1573–1584.

Teh, D., & Rana, T. (2023). The use of internet of things, big data analytics and artificial intelligence for attaining un's sdgs. In *Handbook of big data and analytics in accounting and auditing* (pp. 235–253). Springer.

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Springer.

UN. (2015). Transforming our world: The 2030 agenda for sustainable development. resolution adopted by the general assembly on 25 september 2015, 42809, 1-13. https://doi.org/10.1007/s13398-014-0173-7.2.

UN. (2016). Follow-up and review of the 2030 agenda for sustainable development at the global level, 4.

UN. (2019). *Global sustainable development report 2019: The future is now - science for achieving sustainable development*. United Nations. https://sustainabledevelopment.un.org/content/documents/24797GSDR_report_2019.pdf