# Spatiotemporal Framework for Assessing Urban Performance for Smart Cities

**Jennifer Kim**

Main academic Supervisor: PhD, Zacharie DE GRÈVE, Université de Mons

Partner supervisor: Ramon GRAS ALOMÀ, Aretian Urban Analytics and Design

**A Master Thesis submitted for the Erasmus Mundus Joint Master Degree on Smart Cities and Communities (SMACCs)**

June 2022

University of Mons, Heriot Watt University, International Hellenic University, University of the Basque Country

# Acknowledgements


There is no possible manner through which this thesis could have been produced without the help and support of the incredible people I've been fortunate to be surrounded by throughout not only the last months of the thesis, but throughout this entire two-year epic.

To my family: for always supporting me in all my endeavors.

My fellow SMACCs classmates – with a special shoutout to Christelle, John Paul, Dasha, Danny, and Ice – y'all have been such a joy to study and travel and suffer and laugh with over the last two years, being separated during the last semester was a lot tougher than I expected.

TBO, Diskolaris, and TTT: your personalities extend far beyond the spirit of the game and are the exemplar models for why I continue to chase plastic around the world.

Marc, the best flatmate, for your economist and business rationale grounding me, for our endless conversations about anything and everything, and for our wonderful friendship. Words can't describe just how much I appreciate you.

Ana Gabriela, for being the most incredible cheerleader throughout this whole journey and particularly in the last few days of this thesis. I cannot thank you enough for your hospitality and your generosity and your support over the last seven days.

Joan, the best co-founder I could have stumbled upon, and without whom I genuinely think I would not have a thesis.

Ramon, Jeremy, and Fernando from Aretian, but particularly to Ramon: for his guidance and input throughout the thesis process, as well as validation that all my work chasing after and cleaning data for endless days had value.

Zacharie, for his endless patience, guidance, and (sometimes questionable) trust that I could design and tackle such a different and challenging topic.

To God, to whom be all glory, for His grace and peace, which literally transcends all human understanding, and makes the impossible possible.

# Abstract

In an increasingly urbanized world, 'smart planning' in the form of 'smart governance' has grown to take a critical role in developing smart city strategies. Smart planning requires a comprehensive yet practical urban performance evaluation framework to comprehensively assess and identify the areas of a city most in need of or most benefiting from an intervention. However, evaluating urban performance is challenge, not only due to the ambiguity and complexity of the term, but also because of the difficulty of gathering data structured sufficiently for a comprehensive and continuous assessment. This thesis seeks to propose a methodological framework for evaluating spatiotemporal at multiple spatial and temporal scales for contributing to comprehensive urban evaluation. A database is created for 8 U.S. Metropolitan Statistical Areas over a thirty-year timeframe and a handful of select indicators, incorporating a streamlined methodology for multi-sourcing and estimating missing data. Preliminary analysis using a spatiotemporal lens shows that the combinations of analyses that can be performed using this database may contribute to enriching stakeholders' evaluation and decision-making processes for their locales.

**Keywords:** City science, Geospatial analysis, Spatiotemporal analysis, Database creation, Smart cities, Decision-making, Urban regeneration, Urban performance

Jennifer Kim

June 20, 2022

# List of Tables

# List of Figures

# Nomenclature

**ACS –** American Community Survey

**AHP** – Analytic Hierarchy Process

**API –** Application Programming Interface

**GIS** – Geographic Information Systems

**ICT** – Information and Communications Technology

**KPI** – Key Performance Indicator

**MCDA** – Multi-Criteria Decision Analysis

**MSA** – Metropolitan Statistical Area

**NHGIS –** National Historic Geographic Information System

**SMART** – Specific, Measurable, Assignable, Realistic, Time-Related

**ZCTA** – Zip Code Tabulation Area

# Table of Contents

# 1 Introduction

In the world of urban planning, the concept of 'smart cities' has steadily increased in recent years in city planning vernacular [1]. Although the definition of the term has yet to be fixed [2], the concept of approaching the planning of a city in a S.M.A.R.T. (Specific, Measurable, Assignable, Realistic, and Time-Related [3]), intelligent, or measured manner has gained in popularity. In their haste to become "smart", cities are eager to install new technologies and revise their strategic plans to become more sustainable, technologically complex, and more attractive to existing and potential residents alike. While smart city planning sometimes entails the design of a completely new city from scratch, in most cases, it involves an already existing city which requires assessment to plan about, around, or for it. Prior to developing a strategy for becoming smarter, it is first necessary to evaluate and identify the areas of a city (or, from a perspective of higher scale, the city) most in need of improvement.

Evaluation of urban performance has historically been a challenging field of study, due to the subjective nature of the term. The rise in data processing and storage capabilities with the age of technology has greatly contributed to planners' abilities to collect and process data. However, limitations in data availability and standardization, in addition to behavioral differences across geographies, continue to present large challenges for assessing cities. There exists a multitude of studies measuring indicators associated with a city and many indices which have been constructed by different stakeholders [4]–[9]. However, like the ambiguity of the term "smart city," there also does not exist one given definition or standard for "city performance." These are important to define, or at least to attempt to understand better, to help policymakers and stakeholders identify locations in their jurisdiction that are most in need of investment or have the most opportunity potential for improvement projects.

The main question that this thesis seeks to answer is: in what ways can urban performance be measured and defined using a limited set of common macroeconomic indicators in a spatiotemporal context? Many complex performance indices have been previously developed but are difficult to replicate in different locations and time periods due to challenges associated with data collection, consistency, and standardization [10], [11]. As a result, there is limited literature available on urban analyses which can be reproduced across locations or time periods. Hence, the author was motivated to explore whether a smaller set of common urban indicators could be used as proxy values to evaluate changes in city performance over time.

To address this, a database was constructed for a set of macroeconomic indicators from the U.S. Census and related surveys for eight metropolitan statistical areas (MSAs) in the United States over a thirty-year timeframe in the attempt to identify relationships between the evolutions of the selected indicators. The MSAs selected were evaluated from the Block Group level to the MSA level of aggregation. In addition to evaluating the evolution of these indicators over time and across locations, the indicators were also compared to scores from AARP's Livability Index for possible correlations with community attractiveness as a proxy for urban health.

The contributions of this thesis are as follows:

1) A conceptual framework of smartness and evaluation of urban performance through a spatiotemporal perspective and in the context of decay and regeneration.
2) A deeper understanding of the complexities of urban data availability and analyses.
3) Creation of a consistent database using a detailed methodology for multi-sourcing and estimating missing data.
4) A methodological framework for evaluating spatiotemporal data at multiple spatial and temporal scales for contributing to comprehensive urban evaluation.

The remainder of the thesis is organized as follows: Chapter 2 presents a literature review over relevant theoretical and empirical works to contextualize the reader to concepts of smart cities, decay and regeneration metrics and methodologies, and previous studies in urban analytics, with an extra emphasis on the spatial factor. Next, this theoretical background is linked with the research question and the methodology undertaken for data collection, database construction, and analysis is reviewed in Chapter 3. Chapter 4 provides a description of the case study that the methodology was performed upon, and Chapter 5 summarizes the results from the case study and preliminary analysis. Chapter 6 concludes the work by discussing some practical applications and the limitations of the study, along with some recommendations for future works.

# 2 Literature Review

This chapter reviews the extant literature surrounding the definitions and studies of smart city planning and evaluation of urban performance for regeneration. The theoretical foundations established here provide the context and the driving motivations behind this study. A thorough investigation of the surrounding literature provides a conceptual framework for the reader to understand the complexities surrounding the subjects within the existing literature and the motivations for addressing urban performance from a smart, regeneration-oriented perspective.

## 2.1 Smart Cities

According to the United Nations' 2018 Revision of World Population Prospects, over 55% of the global population resides in urban areas, and this number is expected to increase to nearly 70% by 2050 [12]. This, along with the rise in the term 'Smart City' since the beginning of the 21st century, has prompted many conversations and proposed definitions regarding an ideal concept fit to deal with the logistical and sustainable challenges of an increasingly urban world order. Figure 1 shows the steady increase in online searches on Google for the term 'smart city' since 2004, which experienced an initial peak in 2015, which was followed by another flurry of interest in mid-2020.

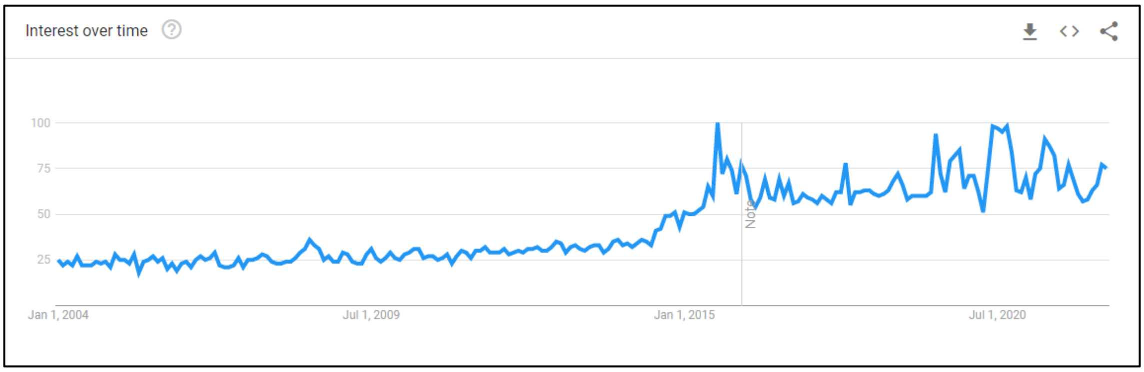


Figure 1: Google Trends search hits between 2004-2022 for 'smart city' [13]

### 2.1.1 Ontology

While the term 'smart city' was originally introduced with a focus on the role of information and communications technology (ICT) being newly implemented in cities, the term was later reframed in a governance-oriented approach to strategic urban planning [2]. Since then, it has been used in a variety of contexts, including corporate marketing, strategic city planning documentation, in reference to technology, and in reference to sustainability – to name a few – yet both definition and usage remain vague and have been referred to as "abstract," "nebulous," and "fuzzy" [2], [14]–[22] and it is interesting to note that the terminology has not been adopted by the United Nations in any formal context [23].

Several attempts have been made to consolidate the existing literature on smart cities [2], [22], [23], reviewing different definitions proposed by stakeholders like Giffinger ("Smart city generally refers to the search and identification of intelligent solutions which allow modern cities to enhance the quality of the services provided to citizens [24]," Caragliu ("A city is smart when investments in human and social capital and traditional (transport) and modern (ICDT) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources, through participatory governance [25]," and Angelidou "Smart cities represent a conceptual urban development model on the basis of the utilization of human, collective, and technological capital for the development of urban agglomerations [26]," among others, to look for commonalities and remark upon differences in definitions and use cases of the term. Despite the absence of consensus on a universal Smart City definition, most of the existing literature tends to converge of on a few key pillars: smart governance, smart people, smart economy, smart living, smart environment, and smart mobility [2], [23], [24]. Figure 2 visualizes these pillars below:

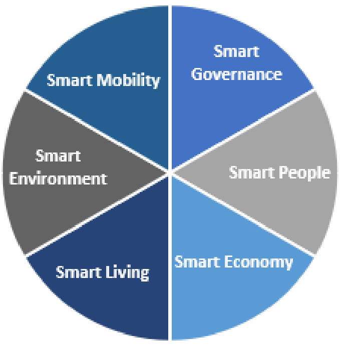


Figure 2: Common smart city pillars

While some stakeholders are prone to associate ‘smart cities’ with technology and ICT, others have been adamantly insistent that the concept should also incorporate strategic planning and sustainability for the environment, the economy, and the society [2], [17]–[20]. Further expanding the coverage of the term, Albino et al. propose that “the label ‘smart city’ should refer to the capacity of clever people to generate clever solutions to urban problems [2].” Shelton, et al., in fact link ‘smart cities’ to ‘urban science,’ claiming that the concept, even if not the term itself, is not new but rather has been around for some time – since the rise of quantitative and computational methodologies used to understand cities in the mid-20th century. According to them, a ‘smart city’ is one that is governed by information and data-driven decisions [16].

### 2.1.2 Governance

Smart governance is one of the six pillars frequently described in the literature on smart cities, not least because the role of governance is inextricably and critically linked to smart city interventions [16]. In the context of the previous section, the role of data-driven governance is particularly important to the strategic planning and decision-making processes of a smart city.

In a different set of ‘smart city planning pillars’ laid out by Koutra et al. [22], one of the key considerations includes “location” – that is, identifying the optimal location to implement new measures or planning investments. This is not only relevant for new construction approaches, but also (and perhaps even more so) is very important for city planners seeking to make improvements upon existing city infrastructure. The location, or the spatial component, is integral, and applicable in a variety of strategic approaches [20]. Evaluating and identifying locations for improvement initiatives results in a doubly smart initiative: smart in the assessment step, and smart in the actual technologies and/or other strategies to be implemented in a way that aligns with the other pillars of the smart city framework.

To assure a data-driven governance model that justifies these decisions, the methodologies and approaches for how to use and apply data effectively must also be addressed. This leads into the next part of the literature review, which further covers existing literature of the concepts and attempts at measuring urban performance.

## 2.2 Urban Performance

The key question to first consider is: what exactly is urban performance, and why do we care? While the concept of urban performance is not standardized and is actually quite nebulous [27], planners and residents alike would agree that it is desirable to live in a well-performing location, whatever that means. Hence, to plan a city smartly that performs well, it is necessary to assess the conditions by which the city can measure its progress or need, whether in absolute terms or in relation to other cities.

One complicating factor is the interchanging use of similar sets of indicators to measure different types of performance associated with cities: sustainability, smart cities, and livability are just some of the other descriptors used in assessing types of urban performance. As a result, within the plethora of assessment frameworks and indices which have been developed to measure various metrics in the context of evaluating a city's performance, it is evident that there is no singular definition that fits everyone's ideal [4], [5], [9], [21], [24], [27]–[29]. However, similar to the observation that the language around smart cities largely centers around six key pillars, so also urban performance may center around particular areas of interest, such as economic output [20], which can help form a baseline for measuring urban performance. Furthermore, as recognized by other researchers such as Caragliu and Del Bo and Balsas, the measurements of concepts such as 'sustainability', 'smartness', and 'livability', all relate to urban performance and can be considered as both constitutional and complementary to it [5], [7]. In fact, these terminologies are often mixed, with some, such as the City Analysis Methodology developed by Leach, et al. [9], emphasizing that these are actually inextricably intertwined. There is no existing 'industry standard' methodology by which one can assess the smartness, or the sustainability, or the performance, or the livability, of a city [28].

As Alberti points out, "urban quality is a combination of tangible and intangible [27]." While perhaps not possible to demarcate specific optimal conditions that are applicable to all cities everywhere, it is of interest to identify measurable performance dimensions which may help explain the relationship between the spatial form of a city and human purposes and values.

### 2.2.1 Measuring Urban Performance

As mentioned previously, to develop a city smartly, or in general, for decision-makers to determine what actions to take to improve their cities, measuring urban performance is critical for the assessment prior to as well as during and in the follow-up of action implementation. Hence, the selection of urban indicators is key to analyzing information about a city in a way that is interpretable and usable by different stakeholders.

Many cities and institutions create their own sets of indicators or indices to evaluate and track the performance of areas of interest. Given the variability in needs and conditions of a city, it is understandable that a city planner might select a curated set of indicators that are more specific and relevant to their particular city's needs and information availability [11]. However, and as a result, there is no streamlined or standardized way of assessing a city's performance which also facilitates comparison with other locations. As a result, unless there is a national or program-level effort to collect and compile data into a singular database, most urban performance monitors operate in silo; otherwise, a very large effort is required for collecting data across different locations, resulting in analyses that occur on a one-off or similarly infrequent basis; for example, studies such as Caragliu and Del Bo's assessment of 94 cities comparing smart city metrics between two time periods [7]. As a result, it is challenging to monitor and evaluate performance over time, resulting in snapshot analyses of discrete conditions rather than exploring the impacts of incrementally changing conditions upon other indicators and their influence upon the dynamics of a city as a whole.

Selecting which indicators to include in a study is challenging; Brown and Kirby acknowledge that selecting components to measure in their urban performance assessment was somewhat in nature [4]; Alberti mentions the same in her discussion on urban sustainability metrics [27]. Corredor-Ochoa also describes the challenge of selecting indicators and additionally for defining the weights to assign to them, as a result of lack of a global consensus [11], an issue which is also described in Lazaroiu and Roscia's "smart city index" assessment methodology [2]. This challenge is further exacerbated by the difficulty of obtaining information that is relevant, measurable, and comparable over locations and time, compounded by each new indicator to be included into the assessment [8], [27].

As a result, some researchers tend to base their data collection from a list of indicators which are either easier to collect, and/or which have previously been analyzed by others [8]. Others justify their indicator selection and the weights assigned through empirical assessments of indicator influence upon proxy overall performance scores using Spatial Autoregressive Local Estimates [7] or through multicriteria decision-making models like the Analytic Hierarchy Process [8] and interview-based weighting methodologies such as the Delphi method [30]. Different stakeholders also have varying opinions regarding the number of indicators that should be included when measuring city performance, since having too many can make the data collection process too complex for cities to practically carry out, while having too few can result in oversimplification of a complex system and lead to biased results [11].

The last thing to note when measuring urban performance is that city performance varies not only between cities (inter-cities), but also in the neighborhoods within themselves (intra-cities). Depending on the scale used for evaluating a city location, contrasting performance results may be observed over different parts of the same city [16], while activities implemented in one part of a city may also exert local impacts upon its surrounding areas [10]. Hence, several researchers emphasize the importance of place-based considerations for analyzing and designing policies. These should take into account not only impacts upon the city as a whole, but also (and especially) for local areas with their location-specific characteristics [4], [7].

Given the cases and challenges discussed above, it is evident why there is no one standardized approach to selecting, weighting, and compositing urban metrics to develop a measure of urban performance. While it is beneficial and generally advisable to follow with historic projects, there is also the possibility of relationships between other indicators remaining undiscovered for lack of innovative investigation. This leaves open a window of opportunity for exploring new combinations of urban indicators in new contexts and frameworks, as explored in this thesis.

### 2.2.2 Spatiotemporal Urban Analytics

Traditionally, many policy analysts and researchers have measured urban performance by capita [31]–[33]. However, focusing only on this type of analysis results in an incomplete picture, since per capita indicators fail to differentiate between general effects of urbanization (i.e., population growth), with local event unique to specific places [32]. They also fail to take under consideration the allometric, or organic, laws that have been shown to be associated with urban dynamics of growth, demonstrating the scalar relationships between many urban indicators and population size [33].

The outputs of a city (such as economic productivity, for example), have been associated with allometric laws that scale with population [32], [34]. Different scaling laws account for differences in urban performance based on the unique characteristics for individual cities; these traits cause outputs that are distinct from other cities that are otherwise comparable in size or geography [33], [35]. Bettencourt, et al., showed that these scaling laws apply for both socioeconomic and infrastructural metrics and remain true over time and space – and that the differences that express the unique features of individual cities also tend to remain consistent as well [32]. To derive these scaling laws and to quantifying the differentiating factors for each city, it is necessary to make observations over time to which changes occur while certain trends remain constant: authors have pointed out that cities experience change very slowly, and the dynamics of urban metrics tend to be dominated by long time scales of at least 30 years [32], [33], [36]. On the other hand, it is also valuable to monitor the changes as they occur over time. As Weaver and Bagley-Sen pointed out in their study [37], having a dynamic view of urban change can help policymakers be better informed and hence more proactive in their decision-making.

It is also important to consider the interconnectivity between geographical spaces and the concept of "proximity," in which the spatial position of a location influences the output of its surrounding locations [35], [38]. Attempting to measure population dynamics is already a complex task by itself; measuring other indicators influenced by population with respect to their spatial characteristics and fitting them to an allometric growth model is a *very* complicated undertaking.

Advancements in geographic information systems (GIS) and computer processing power has helped simplify a lot of these calculations, but most spatiotemporal analyses still focus mainly on evaluating physical changes in a city over time [39] (e.g., urban sprawl [40]). Where they consider soft measurements (or area-allotted quantities, such as vacant housing), they tend to be looked at in still frames, or as comparisons between a limited number of years, primarily for two reasons: 1) data is hard to collect consistently, and 2) boundaries change, complicating efforts to develop apples-to-apples comparisons.

Doraiswamy et al. explain how the complexity of incorporating the three dimensions time, space, and indicator into a database makes it difficult to develop a visual analytics system that is useful and scalable. As a result, "urban-data analysis has often been limited to well-defined questions,… described as confirmatory data analysis, which hampers exploration across datasets essential for understanding trends and potential causal mechanisms [41]." Some solutions and concepts, including some based on Peuquet's "triad framework" for event classification as visualized in Figure 3 below help deal with the complexity of space-time phenomena to the examination of events on an individual level [41], [42].

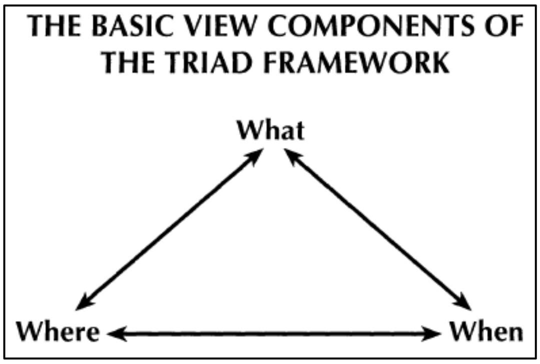


Figure 3: Peuquet's Triad Framework [42]

Although the impacts of time and space are critical to assessment of urban performance, incorporating the multiple dimensions and accounting for the impacts of different aggregations and gradients of time and space present a challenge for urban researchers. This drives the question of whether there are practical analyses and benefits which could be derived from analyzing data sets over smaller increments over time but over longer time frames, while also incorporating an assessment of a city at different scales of aggregation to compare performance of locations within the same area as well as across other cities. Answering these questions may help urban analysts understand, identify, and potentially predict areas that are opportunistic for improvement initiatives.

## 2.3 Urban Decay and Regeneration

For a smart city planner, it is desirable to identify patterns of urban decay to promote regeneration initiatives. Once it is recognized that a specific area is experiencing decay, then improvement initiatives can be selected to optimize the regeneration of the area. Decay is an important term to be able to quantify (or to look for a method of quantifiable measurement) to determine whether a) if an area is in decay and in need of revitalization, and/or b) if an area is in danger of falling into decay, or if the progress of indicators is aligned with a decay pattern. Defining the concept of regeneration is also important since proposals for regeneration can range from the implementation of social programs to tearing down and completely renovating buildings in a neighborhood lot.

### 2.3.1 Definitions

Urban decay doesn't have a singular definition but is commonly identifiable through conditions that remain consistent through the literature: deteriorating infrastructure, higher crime rates, and decreasing real estate value [43], [44]. It is generally described in qualitative terms, based on visible characteristics such as building vacancy and condition, trash accumulation, and of state repairs [45].

The term is mixed often with the concept of "urban blight," which is described as being notable for the qualitative decline in real property conditions [37], [46], along with disinvestment of resources in the area. Some authors, like Cuthbert [47], argue that urban decay, rather than being a preventable condition, is actually a part of the natural lifecycle of a city, and just like an aging human body, areas are inevitably slated to experience a decline in quality as housing stock gets older and people move away, resulting in decreasing market prices and increasing crime rates [44], [48], [49]. For others, it is mainly the result of decisions made by policymakers and investors in favoring some while neglecting other neighborhoods of a city [50]. Once we have identified some areas that seem to be experiencing decline in housing stock (represented either through visual assessments, or sometimes through proxy indicators such as crime rate occurrences, property violations, and rental and sale prices for real estate, it is possible to identify moments from which an area begins to experience decay.

Urban regeneration is a term that incorporates planning initiatives to resolve urban problems and improve an area. This term is often mixed with "urban renewal," "urban development," and "urban revitalization," and although Roberts, et al. would say that they are very different terms, their main differences are mostly in terms of scale and agglomeration used to describe the area being targeted for improvement [51]. In general, however, urban regeneration is "a comprehensive integration of vision and action aimed at resolving the multi-faceted problems of deprived urban areas to improve their economic, physical, social, and environmental conditions [48], [51].

Developing a smart city strategy requires measuring and understanding the performance conditions of a city. These measurements will help identify locations on local and aggregate scales that may be undergoing urban decay and hence opportunistic for regeneration efforts. However, there is still a lack of applicable and comprehensive methodologies to help stakeholders select locations and solutions for successful implementation [52].

### 2.3.2 Identifying Opportunities for Regeneration

There are multiple models available for assessing urban decay and identifying opportunities for regeneration [53], and different methodologies which have been used to quantify and identify decay (and hence opportunities for renewal): using GIS-MCDA tools [54], location based social networks (advantage for identifying unexpected features at a neighborhood scale) [55], quantifying and identifying blight clusters using proxy variables [37].

Indicator-based approaches to identifying opportunities for regeneration have been used quite often [51], [56]. However, there are nearly as many approaches to selecting urban regeneration indicators as there are different studies, leading to inconsistencies in treatment over space and through time [56] and further underlining the difficulty in measuring and identifying locations and activities for planners and policy-makers. Some attempts at development methodologies to address this problems using the following considerations: data availability, geographical specification, time-series prospects, implementation and interpretability [56].

Spatial analysis of urban information is another important part of assessing regeneration opportunities [37], [57]. Within this, it is important to consider scale and the interaction of indicators with spatial features, along with the influence of activities in one space upon its neighboring spaces [55], [58].

### 2.3.3 Decision Tools

Although there is no one-size-fits solution or methodology for selecting a location and initiative to implement for city planners, there are several tools which have been proposed to help facilitate this process.

For example, the Delphi technique is a common technique used to engage with expert stakeholders to identify and quantify key attributes, assigning weights for an analytic hierarchy process or other multicriteria decision analysis evaluation methodology [30], [56], [59], [60]. Some other modeling methods include MCDA, system dynamics, fuzzy cognitive maps, and multi-agent systems, among others [61]. Web-based tools help improve the participatory process from citizens and stakeholders, particularly when combined with geographic information systems (GIS), which helps visualize quantitative information in a more easily consumable manner for interest citizens [54]. The combination of GIS with MCDA frameworks such as agent-based modeling [62], PROMETHEE &

SWOT [63], other multi-criteria comparison methods which consider different spatial scales and renewal scenarios [43], [64], and decision support systems [54] provides a particularly powerful tool for driving decision-making [65].

Even with the many models which have been proposed, there are not so many decision-making models which have been prototyped or actually used in real life [61], with most decisions ultimately being made per business as usual: through vote, or the subjective decisions of a few key decision-makers. In other cases, success stories from specific cities have led to policymakers assuming they can simply copy and paste the same solutions to other cities without considering whether they would be applicable to the circumstances and conditions unique to their own cities [49]. Although it may be more complex and much more difficult, the non-local nature of the similarity among urban trajectories as observed by Bettencourt, et al. further emphasizes the importance of focusing thorough evaluation of specific indicators unique to individual cities and neighborhoods in their unique spatial locations relative to their surroundings [32].

### 2.3.4 Social Impact

Lastly, planning a city smartly not only involves using data-driven and quantifiable evidence to inform and drive decision-making, but also requires consideration for the social impacts that such decisions may have upon residents. Smart cities are not just about improving the economy or improving environmental conditions for a few. It is important to consider the impact on all residents of the city. Equity is a key part of smart city planning (and planning in general) that should not be neglected. Smart city plans should not be conceived as standalone documents but designed around the city's core long-term visions and master plans [66]. Having a multi-scalar framework for evaluating urban indicators will also help stakeholders maintain a holistic perspective of their city and inner communities.

Some unintended consequences of urban regeneration and improvement initiatives may include gentrification and displacement of lower income population, as observed in several case studies [66]–[68]. While smart city urban regeneration initiatives can create new opportunities for innovation and improve service delivery, they may also inadvertently promote spatial inequalities by concentrating resources and infrastructure in select enclaves, further increasing existing social, physical, and economic divides and sometimes even causing displacement of people to city peripheries or other neighbor-hoods [49], [55], [69].

Having these social consequences in mind provides further motivation to perform the most comprehensive analysis possible to pre-identify and prevent or ameliorate situations in which regeneration initiatives only improve some areas or some parts of the population but at a deteriorative cost to others.

# 3 Methodology

This chapter reviews the methodology undertaken in the works of this thesis to develop and construct a database for analysis, visualize the collected data, and to explore the data for quantifiable results.

The methodology of this research was primarily developed under the guidance of Aretian, an urban analytics and design firm specializing in enabling analytical frameworks to derive urban design criteria. Given the scarcity of time-series analyses for urban data in the extant literature, best practices for database construction and imputation were developed with input from several resources, including data science websites and academic papers on imputation.

The study focused on cities within the United States to utilize a consistent data source for the main source of data, the U.S. Census Bureau, to develop an apples-to-apples comparison between indicators. A small selection of variables was chosen to evaluate based on commonly reported indicators. Selecting common urban indicators could facilitate the export of this thesis' methodology to other case studies. With this information, this thesis seeks to address the following research question: "How can a small set of commonly measured urban indicators be used to assess urban performance comprehensively?", using a national livability index as a proxy for city performance, or attractiveness.

To take a SMART approach to assessing urban performance as a precursory step in urban planning – Specific, Measurable, Achievable, Relevant, and Time-Bound, the thesis seeks to determine whether it is possible to streamline the evaluation and decision-making process of a city in identifying opportunities for regeneration by evaluating the assessment of indicators which could be replicated in a SMART manner in other locations.

Figure 4 shows a simplified flow illustrating the primary steps – each of which will be described in detail the following sections – of the applied methodology.

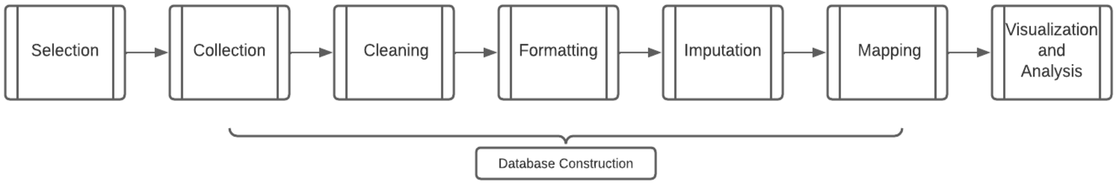


Figure 4: Thesis methodology task flow

The first step, data selection, is built from and builds upon the conceptual framework developed in the literature review. The data collection and cleaning process helps further deepen the readers' understanding of and appreciation for the complexities of urban data availability. The data collection, cleaning, formatting, imputation, and mapping steps describe the details entailing the creation of a consistent database that compiled data from multiple sources and imputed missing data. The visualization and analysis steps demonstrate the methodological framework for evaluating the data at multiple spatial and temporal scales.

## 3.1 Data Selection

To develop a database that covers a range of cities of different performances, this thesis intended to curate a small selection of representative edge cases accompanied by a few other cities in-between. Given limitations in both time and accessibility, the data selected for the database was limited to urban areas in the United States.

To explore urban indicators over multiple spatial scales, Metropolitan Statistical Areas (MSAs) were selected as the reference for urban areas of interest, for each of which the urban indicators of interest were collected at the Block Group level. MSAs are regions that have at least one city and surrounding communities related by social and economic factors [70].

Changes in population between 2010 and 2020 (when the US Decennial Census took place) were used to determine the initial selection of MSAs for the database, since population is a common indicator linked to urban output [32]. Cities experiencing the most positive and negative changes in population were selected, along with a couple of mid-tier cities. Of the 392 number of MSAs in the United States, a limited number of representative cases were selected for the case study, mainly due to the processing power and memory available.

To evaluate the selected indicators in the context of urban performance, a livability index of neighborhoods in the United States was selected as a proxy for urban performance; this index serves as a reference of comparison to which the evolution of the chosen indicators might be correlated.

### 3.1.1 Granularity

The importance of the "unit of analysis" for assessing urban performance was emphasized in the literature review and impressed the need for selecting a level of aggregation for the urban analysis that could be detailed enough to derive more precision but disaggregated enough to avoid introducing biases [33], [34], [71], [72]. On the other hand, there is no fail-safe procedure for defining the boundaries of an urban unit of analysis, and some indicators are more sensitive to spatial restrictions than others [33].

In the United States, the Census Bureau measures and summarizes information at various scales of urban aggregation. Block Groups are the second smallest level of granularity for the United States Census, and generally represent populations of approximately 1500 residents per Block Group. This is also the most detailed level at which census survey data is collected and reported for more variables, since at the block level information may be omitted in the interest of protecting resident privacy. Hence, Block Groups were selected as the base unit of measure; this level of detail permits evaluation of the selected indicators for relative comparisons across both local and aggregated levels. Figure 5 below shows a visualization of the hierarchy of U.S. Census geographic entities.

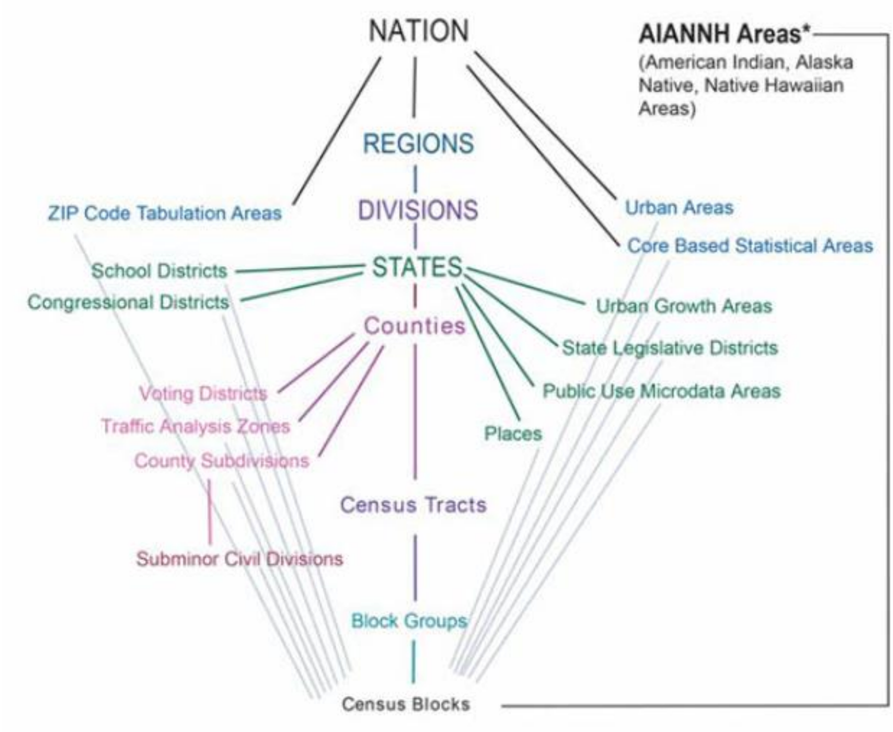


Figure 5: Standard hierarchy of U.S. census geographic entities [73]

Measuring the indicators at a local scale may help identify hot spots and other areas that influence surrounding neighborhoods compared to the agglomerated evaluation of the city. Furthermore, since the U.S. Census does not have the "city" as an explicit unit of data, the database refers to MSAs (Core Based Statistical Areas in the figure above) as the aggregated equivalent of the city level for the analysis. While the database was constructed at the Block Group level, the geographical references to the Census Tract, County, and MSAs were preserved for aggregated analyses.

While the socioeconomic indicators from the census survey are available at the Block Group level, the smallest unit of granularity for the neighborhood livability indices publicly available in the United States is at the zip code level. Figure 5 shows that zip code tabulation areas ZCTAs are not directly aggregated from Block Groups, rather falling only within the national boundary. Hence, it was necessary to create an estimated relation between ZCTAs and Block Groups to ensure a geographically relevant comparison of census variables and livability scores. The U.S. Census Bureau has developed relationship files between ZCTA and Census Tracts, which is one level of granularity higher than that of Block Groups (although it should be noted that not all Census Tracts are evenly aggregated into ZCTAs). Therefore, the Block Groups included in the database were aggregated into their Census Tracts, which were mapped to the corresponding ZCTAs. The livability scores for those ZCTAs were then obtained and assigned to the Block Group levels for incorporation into the database.

### 3.1.2 Time Series

As the literature review revealed that the dynamics of urban metrics are dominated by time scales of approximately 30 years [32], it was desirable to obtain a dataset with data covering this minimum range of time. The database for this case study was therefore constructed based on the availability of the data.

Annual data estimates at the Block Group level are available for most all the selected census indicators for the years between 2010 and 2020. However, to assess city performance over a relative cycle of growth/decay, it is necessary to evaluate data over a longer timeframe. Hence, decennial data was obtained for 1990 and 2000 to supplement the annual data from 2010-2020, using variable mapping to compare like indicators from different surveys. To ensure consistent time intervals of measure in the analysis, data for the missing years in between (in addition to other missing data) was imputed using a mix of interpolation methodologies, which are described in more detail in Section 3.2.4.

### 3.1.3 Key Performance Indicators

As previously mentioned, there is a plethora of indicators utilized in urban performance evaluations which cover a large range of topics and scale. When selecting the key performance indicators (KPIs) for this study, it was essential choose to quantifiable variables that occur frequently in urban performance literature and which are readily available in most, if not all, urban data collection centers. The criteria for the selection was guided by Aretian and other urban studies such as that performed by Mavrič and Bobek [8], who used the following assumptions for the selection of their indicators: objectivity, relevance/measurability/reproducibility, validity, statistical representativeness, comparability/standardization, flexibility, efficiency/performance, accessibility, interaction, consistency and temporal stability.

In particular, it was of interest to select indicators that are consistently collected on a regular basis over a long period of time and for locations available at the Block Group level. The following 11 variables were selected from the information gathered by the U.S. Census Bureau to be included into the study:

- Population

- Civilian labor force
- Unemployed population
- Number of total housing units
- Number of vacant housing units
- Median home value
- Median income per capita
- Median household income
- Median contract rent
- Median gross rent
- Median gross rent as a percentage of household income.

Population, as mentioned previously, is a standard demographic against which many urban metrics are measured, and can be considered a proxy aggregate variable representing the output of various urban mechanisms [32]. Civilian labor force and unemployed population are used to calculate unemployment rate, a classic indicator for economic growth, along with median per capita and household incomes. Total and vacant housing units are used to calculate vacancy rates, which serve as an indicator for attractiveness – that is, more attractive locations should have less vacant houses. This, along with the median home value, gross and contract rents, and the proportion of the rent as a percentage of household income, function as proxies for market appeal.

The assumptions taken in this assessment are 1) Higher rents are indicative of more attractive real estate markets; 2) Unemployment rate and incomes function as proxy measures for economic performance; 3) Population, vacancy rates, and median home values together function as a proxy for urban attractiveness – that is, increases in real estate and rental prices accompanied by population increase are indicative of growth in a city's attractiveness – which can be considered a representation of urban performance.

To verify these assumptions, a livability index was selected as an additional proxy for city performance against which the previously selected socioeconomic indicators could be compared. AARP's Public Policy Institute maintains a livability index that has been curated using over 50 national data sources with 61 indicators to compose livability index scores by zip code for communities in the United States, evaluating categories such as housing, neighborhood, transportation, environment, health, engagement, and opportunity [74]. These scores are publicly available on AARP's Livability Index website tool for the most recent updates (2022) and was previously available for comparison with 2015 scores as well.

## 3.2 Database Construction

The Pareto Principle holds true for data science as well. Data wrangling to construct the database for analysis is a laborious undertaking that requires the use of a variety of resources and tools. Figure 6 depicts a summary of the database construction flow that was followed for this thesis, showing the key sources and outputs as well as the tools used to process the data.

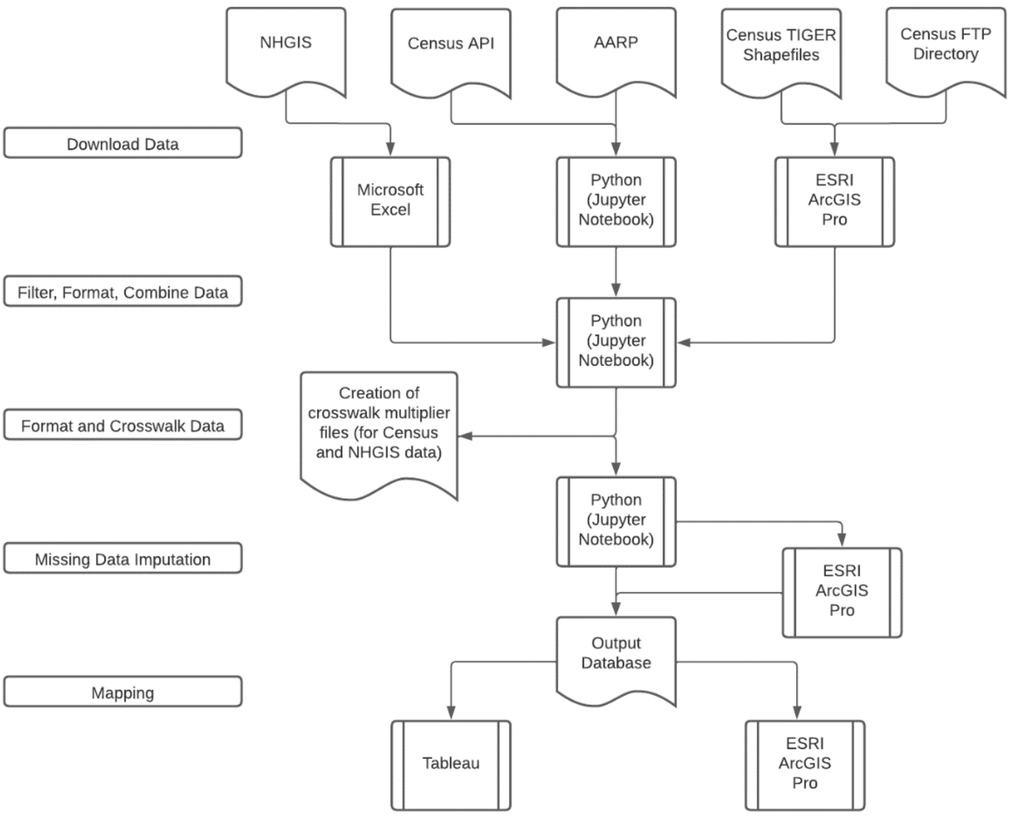


Figure 6: Database construction flow summary chart

First, data was downloaded directly and programmatically from several different sources. Much of this was performed sequentially and not simultaneously as downloaded data sets were explored and gaps found in the downloaded information, which prompted data downloads from other sources. Next, the downloaded data were filtered to extract only the relevant data to be used in the analysis for this thesis, and the resulting data structured and formatted so that the data from different sources could be compatible with each other for comparison. This formatting was not only necessary for the representation of the data but also for the identifiers used to reference the same location over the time series examined. Since the census geographies under examination do not remain consistent from one census to the next, it was necessary to obtain relational references, known as crosswalk data, to transform the data into a standardized set that could be comparable across all years. Once this was performed, the missing data was imputed where possible, before mapping to confirm the completeness of the data, setting up a preliminary visualization for the analysis.

### 3.2.1 Collection

Data was collected and compiled using several different sources and tools. Given the selected time series and granularity, the researcher was faced with the challenge of data paucity in individual resources; hence, it was necessary to seek and gather data from various data types and sources to compile a data set of the desired indicators, aggregation, and time series.

#### KPIs

***U.S. Census Bureau***

The data for the primary database was derived from the U.S. Census Bureau and affiliated government agencies, from which data was downloaded, cleaned, and compiled into a singular database, mainly using Python 3 [75] in Jupyter Notebook. The Bureau's official census database, CensusData.gov [76], was the main data source. This data was supplemented by downloads from the National Historical Geographic Information System (NHGIS) [77] and from geodatabase file extracts from a separate repository of data maintained by the Census Bureau [78]. When selecting a livability index to use as a proxy, the salient rating institutions for measuring livability in the United States were AARP's Livability Index and Livability.com. AARP's Livability Index was selected to be used as the primary proxy against which the selected urban indicators would be compared.

The U.S. Census API at the time of the study only contained data available at the Block Group level for 2013 to 2020. This data was downloaded using the opensource Python package CensusData [79]. The data downloaded was sourced from the American Community (ACS) 5-Year Survey, selected due to the relative reliability, consistency, and availability of the variables desired. However, because the variable tables occasionally change from one year to the next, the variable tables for each year also needed to be reviewed to ensure consistency in variable nomenclature from one data batch to the next.

Although the U.S. Census Bureau website states that ACS 5-Year Survey Data is available at the Block Group level from 2009 to 2020, the API only provided data from 2013 onwards. The Census Bureau's data dissemination platform, data.census.gov, was helpful for discovering and downloading smaller sets of data, but not practical for downloading specific sets of variables for specific geographies at the Block Group level, both due to format of data tables, as well as due to the bulk size of the data downloads.

The U.S. Census Bureau's FTP site contains geodatabases for annual ACS survey results on the Block Group level for several years [80]. These were used to supplement the information that was not available through the Census API. More specifically, geodatabases with ACS 5-Year Survey data at the Block Group level was downloaded for the relevant states (for some years, the data was divided by state; for other years, the data was agglomerated into a single dataset with multiple relational tables containing values for the different variables in the survey.) Data was downloaded for 2010, 2011, and 2012 to supplement the data directly downloaded through the Census API.

The geodatabase files downloaded from the Census Bureau's FTP site were opened and handled directly in ESRI ArcGIS Pro [81]. KPIs for the desired Block Groups were copied from the shapefile tables and stored in Excel to be compiled with the other downloaded data in Python. However, upon closer inspection, some of the 2010 and 2011 data were missing from the geodatabase files, and hence were populated with data obtained from the NHGIS database where possible and prepped for imputation where unavailable.

***National Historic Geographic Information System***

For the pending information for 2010 and 2011, along with older data, it was necessary to perform data queries for the desired KPIs from the NHGIS online database [77]. The data was available for download in Excel format at the Block Group level but contained in sets of data by state; hence, it was necessary to download multiple files of data which was then cleaned and processed to filter only the relevant data for use. The KPIs obtained from the NHGIS database was sources from various surveys – for the 2010 and 2011 data, information from the ACS 5-Year Survey was accessible. For 1990 and 2000, some of the KPIs were available from the 100-percent data summary files (SF1), and others were only available from sampled data summary files (SF3). Furthermore, because variables change in name and appearance across different types of surveys and years, it was necessary to construct some of the KPIs (for example: data for civilian labor force and unemployed population for the year 1990 were constructed through a calculation and summation of labor force information that was only available in demographic breakdowns) and to map others. Appendix A shows a full table of the indicators downloaded with the breakdown of the associated variable name, survey, reference table, and the source method from which the data was obtained.

***AARP***

The most recently updated AARP Livability Index scores (2022) and for 2015 were extracted from the website after mapping Block Groups to their respective zip codes using the Census Bureau's relationship file (described in Geographies below). Figure 7 and Figure 8 show the AARP Livability Score results which can be viewed from AARP's online tool [82].

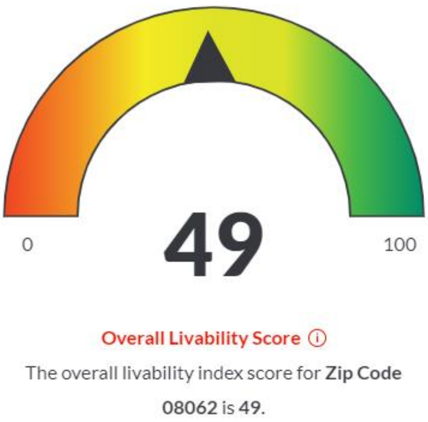


Figure 7: Overall Livability Score for 08062

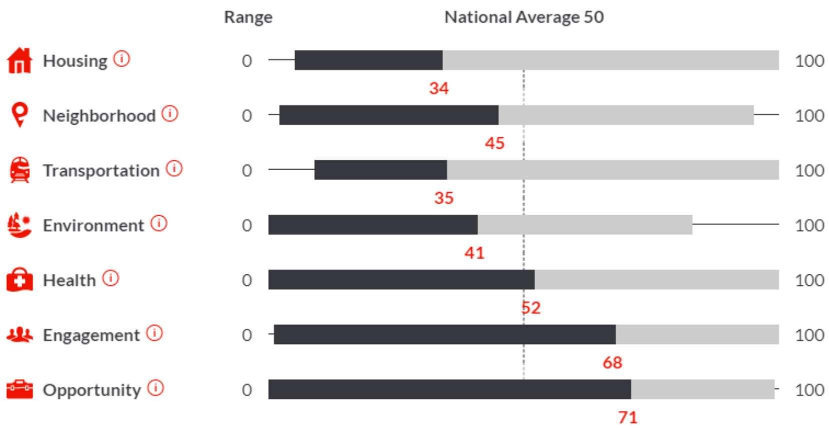


Figure 8: Categories scores comprising the overall livability score for 08062

Given that U.S. Census Block Groups do not fit perfectly within the boundaries of zip code tract areas, there were a number of Block Groups which had to be manually reviewed on a map overlaying Block Group boundaries and ZCTA boundaries to evaluate and determine into which zip code (if any) a given Block Group could be assigned. From this mapping exercise, a complete list of zip codes associated with the Block Groups could be compiled and the AARP Livability Index scores extracted for them. In addition to the composite score, the individual category scores were also downloaded. There were a few cases found in which zip codes were associated with other locations due to the lack of location or size of the area under consideration. For those 22 cases, the original zip code was manually input into AARP's website tool and the corresponding zip code value recorded and replaced in the data query.

### Geographies

The U.S. Census Bureau also has TIGER/Line Shapefiles, which are extracts of geographic and cartographic information from the Census Bureau's Master Address File (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER) Database (MTDB) [83]. The 2020 shapefiles were available for download by state and were cleaned to limit the shape of the data to those areas within the geographical area of the study. These

TIGER/Line Shapefiles for 2020 ZCTAs were also downloaded to be joined with the 2020 ZCTA to Census Tract Relationship File [84], which was used to relate the AARP Livability Index scores (provided at the ZCTA level) to Census Tracts, then further extrapolated to the Block Groups in the case study database.

Finally, the geographical boundaries of the census units change to varying degrees with every decennial census. As a result, it was necessary to crosswalk the older data to a reference year's geography in the intent of creating an apples-to-apples comparison over the time series. Crosswalk files were available at the Census Block level from NHGIS – for 1990-2010, 2000-2010, and 2010-2020 crosswalks [77]. These files contain the information relating Blocks from the older year (referred to as the source) to the more recent ones (referred to as the target). They also contain interpolation weights which have been derived from advanced models, that indicate the proportion of a source area's characteristics which ought to be allocated to the target area [85], which were applied to the KPIs associated with each location. Given that the crosswalks were not available at the Block Group level for all the years to be transformed, additional steps were required to apply the conversion factors from the geographical boundaries of a Block Group from the source year to the target year. These are described in more detail in the sections below.

### 3.2.2 Cleaning

Once all the data was selected, discovered, and downloaded, there were several steps that were iterated through for all pieces of data, to clean and prepare it for formatting. Most of the data was treated as a separate dataset for each year, given the different sources and tables involved.

First, because datasets were only available to be downloaded for the entirety of the United States or by state (which also required a geographic indexer that correlated all the MSAs of the study database with the states in which corresponding Block Groups are located), each downloaded set had to be cleaned to extract only the desired KPIs associated with the geographies of interest. This was done primarily to reduce the processing demand, given the size of the data download.

For example, for the data extracted from the geodatabases downloaded from the U.S. Census Bureau, ESRI ModelBuilder was used create a workflow iterating through the feature layers and tables to extract only the information associated with the Block Groups within the selection of MSAs and for the chosen KPIs, as visualized in Figure 9 below.

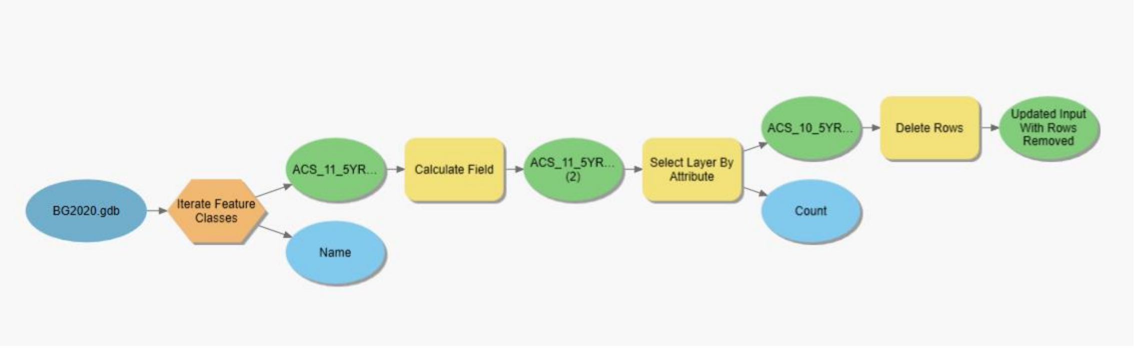


Figure 9: ArcGIS Pro ModelBuilder data extracting flow design

For data downloaded into Excel format, a combination of manual manipulation in Excel and Python coding was deployed to extract files from downloaded zip folders, concatenate multiple Excels into single sheets, and then filtered for data to be kept. The datasets that were manually downloaded were constructed in Excel to match the same format as that established for the 2013-2020 data in Python. To filter out the non-geographically relevant data, a new column field was created to identify the state and county codes for those relevant Block Groups, and all those outside of the desired geography were removed from the database.

Next, several fields, mainly geographic reference fields not already included such as MSA, state, county, zip code, etc., were constructed and joined to each of the datasets.

Since different datasets were obtained from different sources, there were many different formats in data type, fields included, and even field name. Every set of downloaded data was evaluated and treated to ensure consistency across years, in anticipation of being combined into one singular data frame. All data was constructed using Excel or Python to ensure that all data sets have the same fields.

Finally, the last stage of the cleaning process was to evaluate the data for missing, null, or negative values. Null values were represented differently across the different downloads; in some years, null values were represented as blank (<Null>); in others, these were represented by the value -666666666; lastly, there were several zero values for indicator fields, which were not clarified as to whether or they were null. The initial assumption made was that there are no zero value indicators; hence, everything that was marked as <Null>, <NaN>, a negative number, or zero, was set to a Null value.

### 3.2.3 Formatting

Following the data collection and initial cleaning, data wrangling was performed to format the datasets into analogous structures which could then be merged into a singular data frame.

#### Structure

First, variable mapping was performed over all the datasets to ensure the same variables were being recorded for the same locations over each year. Next, each dataset was formatted into tables with data types consistent across years.

#### Crosswalks

As mentioned in an earlier section, the physical boundaries of the urban units change with every census year. As a result, it was necessary to crosswalk each of the datasets and translate the data into a new format standardized to a selected year (in this case, 2020, as it was the most recent year of available data as well as the latest census update). Although there should not be major changes, they are significant enough that crosswalk files are produced to help researchers convert their source data to a target year to help stabilize the comparability of different sets of data. The data crosswalking process is visually represented in Figure 10 below.

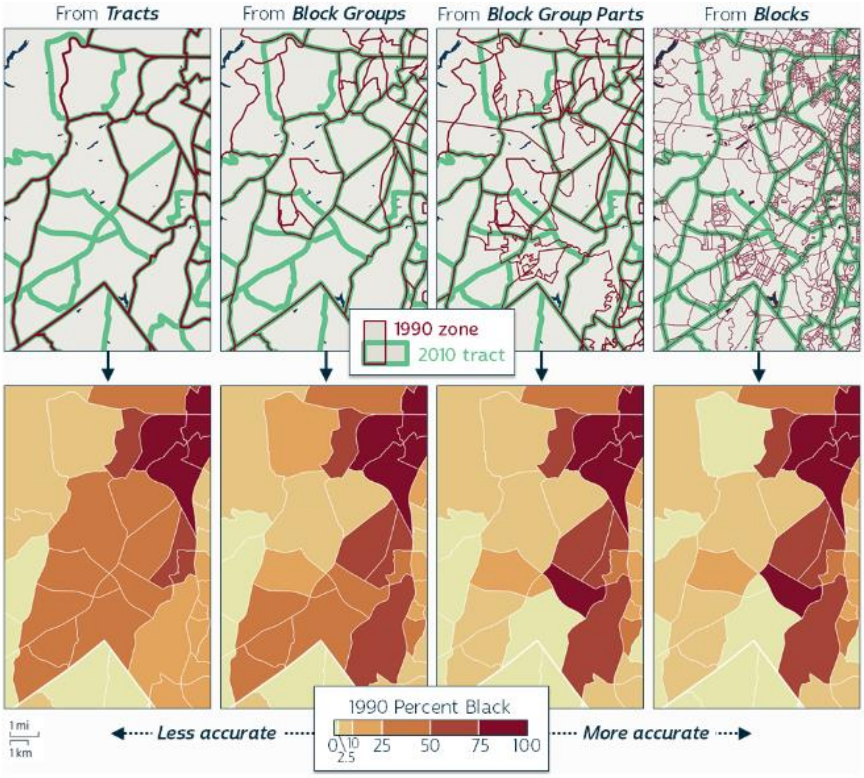


Figure 10: Visual representation of crosswalk mapping [85]

The NHGIS provides crosswalk files for 1990 to 2010, 2000 to 2010, 2010 to 2020, and 2020 to 2010. However, the geographic level availability is restricted for earlier years; thus, only Block to Block crosswalks are available for all the time frames of interest. The crosswalk files also include interpolation weights (ranging between 0 and 1) that approximate the distribution of the proportion of a source zone's population and housing units located at the intersection between a source unit (from one year) and a target unit (from the second year) [85].The NHGIS additionally provides information on how to use crosswalks; while it was attempted to follow the suggested methodology, some additional steps were required to be developed and applied as well.

To crosswalk all the data sets to the 2020 geographic boundaries, the crosswalk files were first downloaded for 1990 to 2010, 2000 to 2010, and 2010 to 2020. As these were stored in files by state, it was necessary to download all the Block level data for all the states into which the case study MSAs extend. These were then filtered and reduced only the crosswalk data relevant to the locations in the case study database. These were merged to form multipliers to transform the 1990, 2000, and 2010 data series. Since crosswalk files were available for mapping 1990 data to 2010 boundaries, then 2010 data for 2020 boundaries, the 1990-2010 and 2010-2020 crosswalks were merged in using an inner join in Python Pandas [86] to create a 1990 to 2020 crosswalk. The same was also performed on the 2000 to 2020 crosswalks.

It was observed that some Blocks in earlier source years (for example, in 1990) were not listed in target years (for example, in the 2010 crosswalk), but would show up again in a later crosswalk set (for example, not be existent in 2010 but have been added back into the 2020 crosswalk). As a result, there were several situations in which there were multiple Blocks from a source year discarded or mapped to one single Block in the target year, and vice versa as well. To address this, interpolation weights of NaN were replaced with 1 and the interpolation weight of one crosswalk multiplied with the interpolation weight assigned to the joined Block.

To transform the data, the Block Group reference was extracted from the geographic identities of the Blocks in the crosswalks, and then the Block Groups grouped by maximum multiplier weight. This was based on the assumption that the unit of area with the greatest weight is the one with the most overlap, and therefore the most likely to actually fall within the Block Group it was being mapped to. Once the crosswalk data was grouped, it was then merged with the dataset for transformation with the interpolation weights multiplied with the count-based indicators. This was then grouped once more (since there multiple Block Groups may have been mapped to the same target Block Group). In the grouping, the Block Groups were again aggregated by maximum values for each of the indicators. This was performed three times: for the 1990 data set (using 1990-2010 and 2010-2020 merged crosswalks), the 2000 data set (using 2000-2010 and 2010-2020 merged crosswalks) and the 2010-2019 data series (using 2010-2020 crosswalks).

### ZCTA to Block Group

The AARP Livability Index data is only available at the Zip Code level of granularity; hence, this data needed to be formatted to fit into the Block Group level of aggregation of the database. Since ZCTA to Block Group relationship files were not available for 2020 geographies, the ZCTA to Census Tracts Relationship File was used instead; since Block Groups are nestled into Census Tracts, it is relatively easy to map the Livability Index values which had been mapped to Census Tracts back to the Zip Code level.

### 3.2.4 Imputation

Once all the data was cleaned, formatted, and mapped into a comparable data structure, the last step remaining was to impute for missing values. For this database, there were three main types of missing values: 1) civilian labor force and unemployment population for 2011 (as this data was not available at the Block Group level in any of the data sources), 2) all data for the years between 1991-1999 and 2001-2009, and 3) missing values randomly distributed throughout locations, years, and KPIs.

#### Strategy

To approach this task, several different methods for imputation of missing data in time series were evaluated. Some studies replace missing data through machine learning algorithms using nearest neighbor correlations [87] or pattern sequence forecasting [88]. Li and Revesz also examine various methodologies for interpolating spatiotemporal data [89]. On the other hand, simpler ways of dealing with missing data values use SciPy's imputation functions [90] to fill missing values through forward fill, backwards fill, arithmetic mean or median, or by completely dropping those locations with missing information [91], [92].

Given the interest of the study to provide a streamlined methodology for constructing the database, complex machine learning algorithms for estimating missing values were not considered for the scope of this task. Regression models were also not considered due to the time sensitivity of the data, lack of data points, and lack of knowledge of the proximity influence of Block Groups upon their neighbors. Hence, the assumption that Block Groups experience change independently over time was implicitly included in determining the approach for filling in missing values.

Simple imputation methods filling constant values, whether forward, backwards, average values or other, are less representative of a time series, given that urban performance is dynamic and changes over time. Since all the Block Groups had at least one missing value amongst the KPIs under evaluation for the 30-year time range, it was infeasible to drop missing data locations from the analysis.

First, all data was compiled into one data frame and inspected to see how much and which data was missing across the 30-year time series. Block Groups that did not appear consistently across all the years were removed from the database, as well as rows completely missing data for the Census-based indicators. For the purpose of simplicity and data contiguity, it was assumed that zero values were representative of null cases – it should, however, be noted that the zero values may be reflective of the census' attempt to preserve individual privacy, and/or may not actually have been measured or recorded for that year.

### Missing KPIs for 2011

Linear interpolation was performed to fill in the missing civilian labor force and unemployment values in 2010. Linear interpolation was selected based on the availability of the existing data, and since linear interpolation is one of the most straightforward methods used for imputing missing data for time series [93]. Interpolation using preexisting functions in packages such as SciPy were originally attempted but encountered problems resulting from randomly missing null values. As a result, it was necessary to write a function in Python that would examine each available year of data from 2011 and search for two valid data points for a given Block Group and KPI with which a linear trendline could be constructed. If and when these points were found and the linear equation derived using a NumPy polyfit function [94], fill values were calculated and placed in for 2010. If there were any Block Groups that did not have at least two years of data points for an indicator between 2011 and 2020, then the 2010 value was left blank.

### Missing Data Between Decennial Censuses

Following this, gaps in the "edge" years (1990, 2000, 2010, and 2020) were padded using spatial imputation. The data for these years was exported from Python into Excel and imported into ArcGIS Pro, where it was mapped, then "Fill Missing Values" function performed using ArcGIS Pro's Data Engineering toolbox [81]. Spatial imputation was performed using the median value of neighboring Block Group features sharing or overlapping in boundaries. Spatial imputation was selected since the missing data did not have sufficient temporal neighbors with which to back- or forward-guess the value. Since the datasets were then examined in their yearly cross-section, the only surrounding pieces of information which could be relevant was the spatial neighbors.

The median value was selected over other options such as average, zero, K-nearest neighbor to avoid potential outlier cases in the surrounding Block Groups as well as trying to ensure that proximity conditions were respected. Even then, some Block Group values were left empty due to insufficient or no bordering neighbors.

Next, to fil the gaps between the decennial censuses which were conducted in 1990, 2000, and 2010, linear interpolation using a modification of the function created for the previous interpolation task was conducted for each Block Group for the years in between each census year. Given the large gaps in data, each census year pair was treated separately for the linear interpolation. That is, linear interpolation was performed for 1991-1999 using the data points from 1990 and 2000, then linear interpolation performed for 2001-2009 using data points from 2000 and 2010.

#### Randomly Distributed Missing Data Imputation

Following completion of the previous steps, a cubic spline interpolation function from the SciPy library [90] was utilized to impute missing data points for the years between 2010 and 2020. Linear interpolation was initially considered for implementation to remain consistent with the previous imputation steps. However, in the interest of creating a curve that is more responsive to changes over time, and also given that the missing data for this range of data occur more randomly instead of consistently across the entire dataset (as for 2001-2009 and 1991-1991), there was better distribution of data available for apply the spline function for interpolation.

### 3.2.5 Mapping

Once all the data was finally completed, it was joined to its geospatial features and mapped in Tableau [95] and in ArcGIS Pro [81]. Mapping in either program provides a quick first overview of the completeness of the data, and to verify whether there were any large areas missing data in the database. Once this is confirmed, then mapping becomes the first step in the exploratory data analysis and preliminary assessment.

## 3.3 Data Analysis and Visualization

Extensive exploratory data analysis was conducted upon the output database. Given the quantity and multi-dimensionality of the data collected, there were many relationships and trends to explore. First, some descriptive statistics were collected for the dataset, followed by examinations for trends in time series, then in cross-sectional relationships. In this third section, the cross-sectional relationships were explored not only between the Census KPIs, but also in comparison to the Livability Index scores.

### 3.3.1 Descriptive Statistics

First, general descriptive statistics were obtained for the data. Descriptive statistics are useful for summarizing the data to look for salient observations or trends. General statistics were calculated over the database as a whole, and also over the most recent available year of data, in order to have an overview over both the time series and across locations in a given year. Asides from standard statistics such as average, mean, max, min, the information was also summarized using a medley of summary tools such as box and whisker plots, histograms, violin plots, and the likes. The descriptive statistics were obtained and visualized through a combination of Python libraries [75], prepared through Tableau Prep Builder [96], and visualized in Tableau [95].

### 3.3.2 Time Series Trends

Next, the data was examined in Tableau over the time series in the search for trends over time which could be identified for the various KPIs in the database. These were evaluated on a global trend as well as for the individual MSAs as well. Gradients for trend lines were also of interest to see if there were any emergent patterns in the data. Although of interest, methods for forecasting were not considered in the scope of this thesis. However, correlations between variables were explored by examining the evolution of indicators and their changing slope with respect to each other and seeing whether a pattern could be identified. This was of particular interest when looking for a temporal relationship between changes in median household income and gross rent.

### 3.3.3 Cross-Sectional Relationships

Following the longitudinal assessment of the data, a latitudinal – or cross-sectional – assessment was carried out over the data, using multiple years for analysis: 1990, 2000, 2010, 2020, and 2015 (since the AARP Livability Index scores are associated with this year). The cross-sectional analysis examines the variables in relationship with each other and across cities, over different urban units of aggregation. N by N matrices of the KPIs were created to provide a quick look at the correlations between variables. Logs of the indicators were also graphed against the logarithmic value of population to explore whether the urban laws of scaling are applicable for these indicators. Lastly, the KPIs from the Census were plotted against the different categories of the Livability Index to check for any possible correlations. While most of the analysis was mainly performed in Tableau, ArcGIS was also used to evaluate the data for clustering and outlier analyses using Anselin Local Moran I's model [81] to identify concentrations of high values, concentrations of low values, and spatial outliers [97].

# 4 Case Study

This chapter presents the case study crafted for assessing city performance in this thesis. As described in the Methodology, to develop a conceptual understanding of urban performance, it was necessary to prepare a database with a range of cities in terms of their subjectively known performance while limiting the study to a manageable quantity of data.

Table 1 below shows the list of MSAs selected to evaluate in the case study. The number of counties, the 2020 census population count, and changes from the previous census are shown, as well as the approximate area of each and the number of Block Groups which will be evaluated for each MSA. It should be noted that MSAs can differ quite significantly in both population and land area size; a note was made to keep this in consideration when evaluating the data.

Table 1: Selection of Metropolitan Statistical Areas for the Case Study

| MSA | State(s) | 2020 Census (Population) | % Change from 2010 | Area (km2) | # Counties | # Block Groups |
|---|---|---|---|---|---|---|
| Charleston | WV | 258859 | -7% | 6910 | 5 | 240 |
| Delaware Valley | DE, MD, NJ, PA | 6245051 | 5% | 12614 | 11 | 4306 |
| Flint | MI | 406211 | -5% | 1682 | 1 | 373 |
| Greater Akron | OH | 702219 | 0% | 2394 | 2 | 553 |
| Greater Austin | TX | 2283371 | 33% | 11085 | 5 | 967 |
| Greater Boston | MA | 978529 | 7% | 11684 | 7 | 3418 |
| Greater Cleveland | OH | 2088251 | 1% | 10307 | 5 | 1694 |
| Metro Detroit | MI | 4392041 | 2% | 10970 | 6 | 3698 |

As seen in the table, the Greater Austin area in Texas experienced a huge amount of population growth between 2010 and 2020. Given this statistic, the city was selected with the expectation that it would demonstrate an optimal example of urban performance. Charleston, West Virginia, ranked on the other end of the population growth scale, and recorded the greatest decrease in population between 2010 and 2020. However, given the relatively small size of the MSA, other larger cities were selected to include into the comparison for that end of the analysis. Hence, Metro Detroit was selected as a counterpoint to Austin, given that Detroit has been called a city that failed [45]. Interestingly, Metro Detroit experienced a slight positive change from the 2010 census, but is located adjacent to the MSA of Flint, which had the second highest population loss between the 2010 and 2020 census.

Akron and Greater Cleveland are like Flint and Greater Detroit in terms of population change and proximity with respect to each other; hence, these were also included into the database in the interest of viewing if and how neighboring MSAs might impact each other. The database was rounded out with other large MSAs that experienced a medium change in population – Delaware Valley, which is centered around the city of Philadelphia, and Greater Boston. The selection of MSAs for the case study are mapped in Figure 11 below.

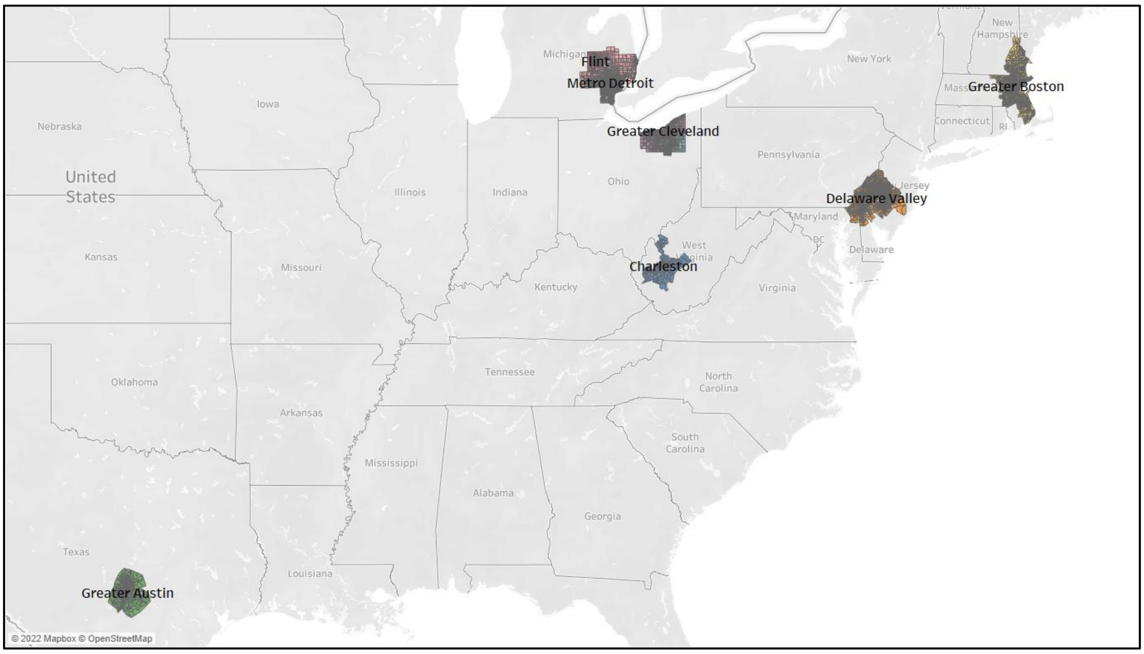


Figure 11: Metropolitan Statistical Areas included in case study database

With the selection of indicators described in the methodology above, the case study aims to provide an example of the methodology and the frameworks taken for novel spatiotemporal urban analytics. Hence, the goals of the case study are as follows:

- Walk through the construction of a consistent database spanning time and geospatial scale.
- Evaluate the resulting spatiotemporal dataset through the lens of the methodological framework assessing information at different spatial scales, time series, and against different KPIs.
- Attempt to draw some mathematical correlations between the time series evolution of urban performance trajectory.

# 5 Results

This chapter reviews the experience of building the case study database, in addition to discussing some preliminary results from the exploratory data analysis and correlation analyses conducted following completion of the database.

## 5.1 New Database and Missing Value Imputation

Using the methodology described above, a database was created for 8 MSAs in the United States, collecting and imputing urban indicators from the U.S. Census Bureau for a 30-year time series. This database was also joined to the 2020 geodatabase features for the Block Groups in the selection of MSAs, as well as to the 2015 and 2022 livability index scores taken from the AARP Public Policy Institute.

The original selection of MSAs encompassed a total of 15,947 Block Groups according to the 2020 geospatial boundaries. After crosswalking all data from 1990 onwards and mapping to the 2020 geographies, 15,514 Block Groups, or 97% of the original set, were left with consistent data over all the years of the dataset. The resultant data table of KPIs consists of 480,934 rows and 22 columns of data, of which over 99.9% of all columns have existing or imputed values. Figure 12 shows one of the visualizations of population distribution for the year 2015 in the Greater Austin MSA.

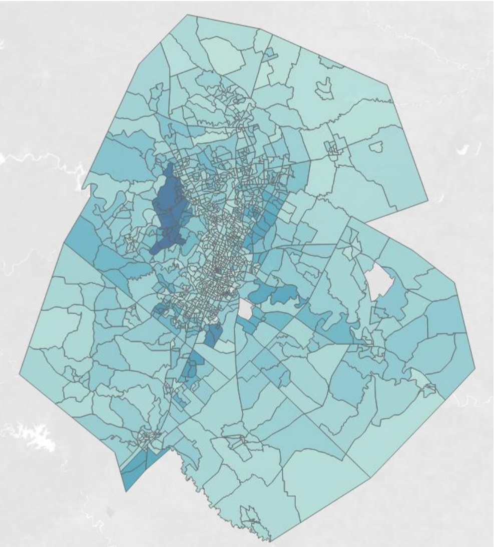

Figure 12: Population distribution in Austin for the year 2015 in Austin

## 5.2 Descriptive Statistics

Descriptive statistics show the preliminary distribution of the database that was constructed in the case study. Table 2 shows the description of the KPIs in the database across all years. Great variation is evident across years and locations. One thing salient observation is that although Block Groups were generally defined to contain approximately 1500 residents, there are some entries in the database with much higher population-associated counts; this is indicative of outliers.

Table 2: Statistical description of case study database

| | Population | Median Household Income | Per Capita Income | Total Civilian Labor Force | Unemployed Population |
|---|---|---|---|---|---|
| **Count** | 480934 | 480904 | 480934 | 480924 | 480914 |
| **Mean** | 1640.3 | 62365.4 | 29615.5 | 863.4 | 60.3 |
| **Std Dev** | 997.2 | 33390.1 | 16867.6 | 562.7 | 51.4 |
| **Min** | 1 | 2499 | 199 | 0 | 0 |
| **25%** | 982.2 | 39063 | 18581.33 | 489.4 | 26.6 |
| **50%** | 1372.3 | 55527 | 25908 | 714 | 47 |
| **75%** | 2027 | 78210.1 | 36079 | 1091.6 | 78.6 |
| **Max** | 19877 | 250001 | 361057 | 11185 | 1500 |

| | Total Housing Units | Vacant Housing Units | Median Contract Rent | Median Gross Rent | Median Gross Rent as Percentage of Household Income | Median Home Value |
|---|---|---|---|---|---|---|
| **Count** | 480904 | 480883 | 480733 | 480792 | 480824 | 480481 |
| **Mean** | 676.3 | 56.1 | 788.5 | 937.8 | 30.9 | 207956.9 |
| **Std Dev** | 397.1 | 61.7 | 393.9 | 414.5 | 9.9 | 159783.4 |
| **Min** | 0 | 0 | 0 | 0 | 0 | 0 |
| **25%** | 410.6 | 20.3 | 525.4 | 647 | 24.45 | 101100 |
| **50%** | 570.4 | 38 | 702.8 | 853 | 29.43 | 165220 |
| **75%** | 835.5 | 70 | 954 | 1130 | 36.18 | 266840 |
| **Max** | 8094 | 1461 | 6569 | 5237 | 318.2 | 2000001 |

By examining a histogram of the distribution of population, for example, it is evident that while most of the data related to population is within or around the expected quantities, the distribution of population quantity across different Block Groups and years does vary. Figure 13 shows the distribution of this data.

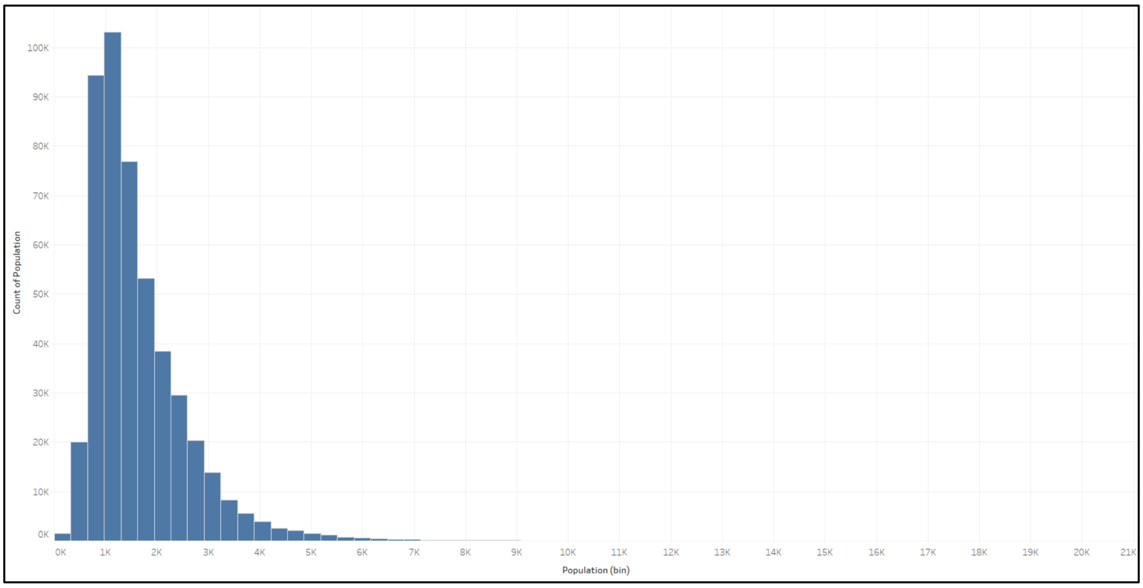


Figure 13: Historigram showing population distribution in the case study database

Other types of descriptive statistics include bar graphs, which can be used to show the change in summarized data over the time series of the database. It is interesting to note in Figure 14 below how population seems to be growing consistently but experiences a large drop in 2020. This may be indicative of impacts due to the COVID-19 pandemic which impacted moving and census survey responses but could also be an indication of something to check in the database and the construction process. On the other hand, the other indicators in the bar graph don't seem to experience as big differences in 2020 compared to in previous years.

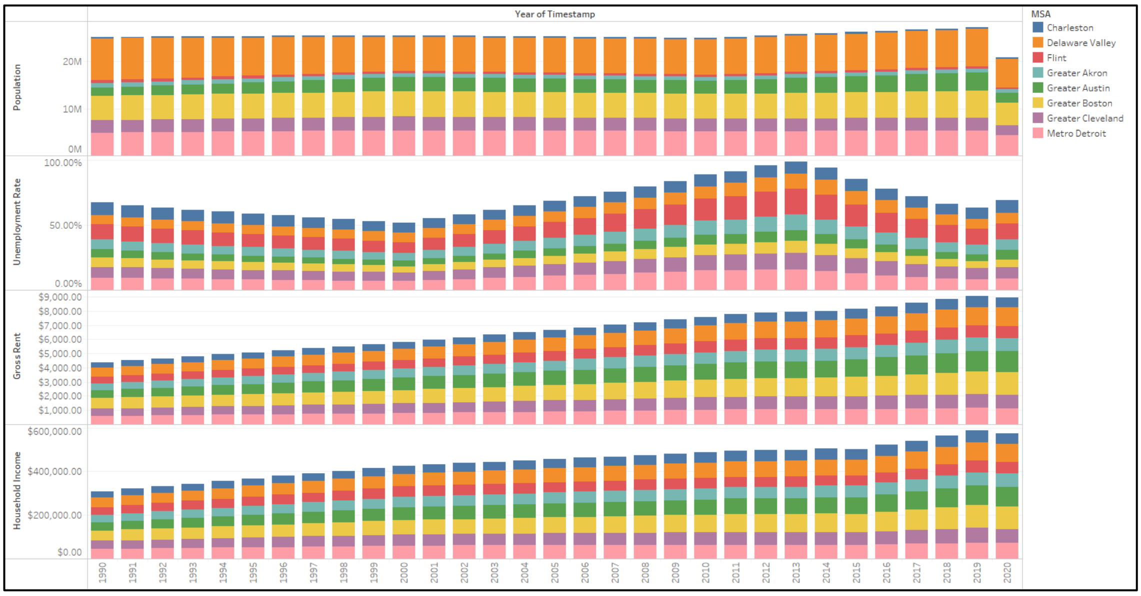


Figure 14: Bar graphs for a sampling of KPIs and their change over time per MSA

For additional exploration of the dataset, there is a medley of summary tools that can provide more information about the dataset through additional tools such as scatter plots, box and whisker plots, histograms, violin plots, and the likes. The following figures show a sampling of these, showing some of the relationships which can be explored visually and investigated in further detail to derive quantitative observations.

Figure 15 demonstrates one example of using scatter plots to get an idea of general truths, such as the observation that bigger MSAs (e.g., Delaware Valley) tend to have bigger economic output, as visible in the distribution of the orange dots towards high values.

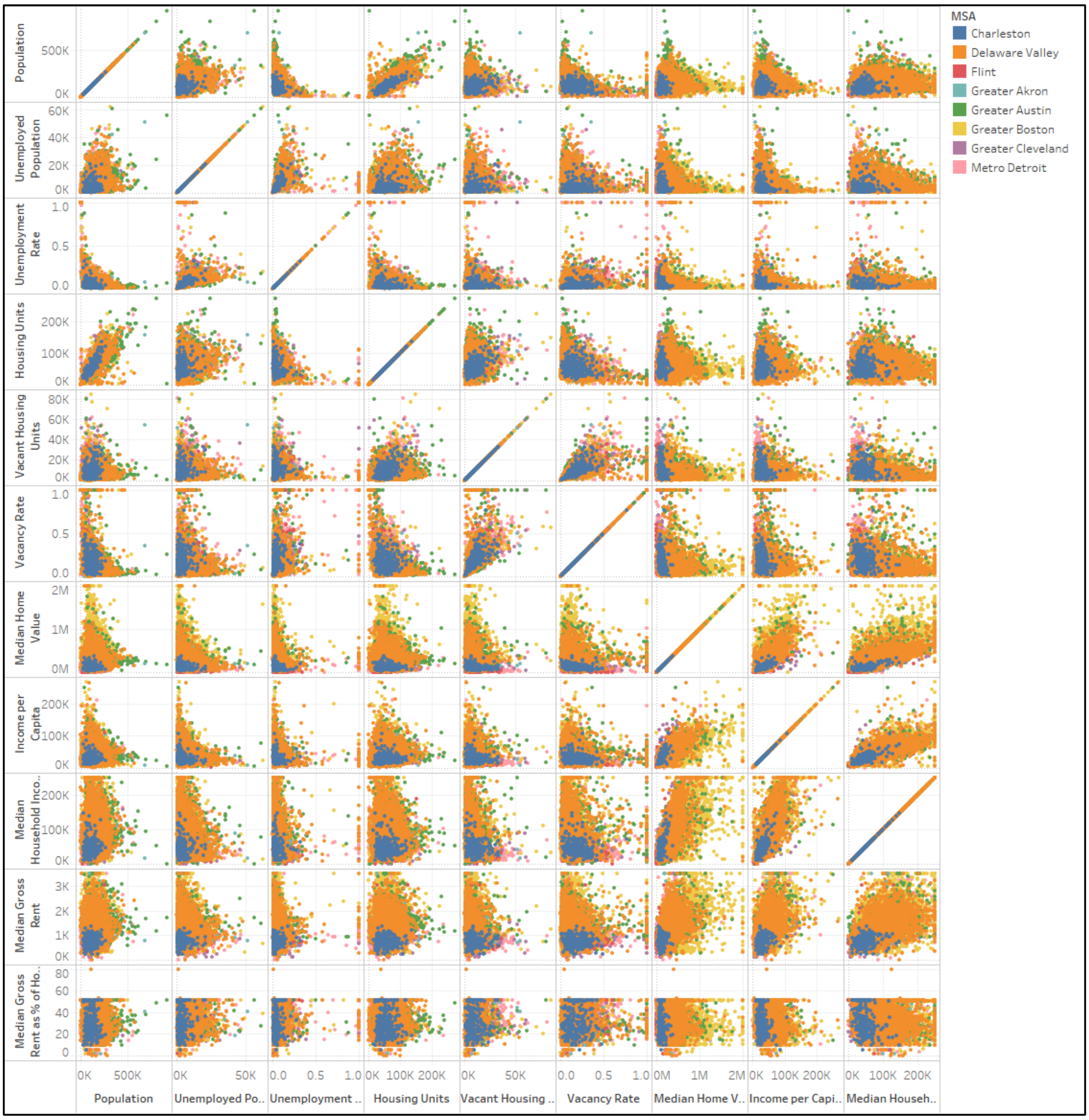


Figure 15: Scatter plots of KPIs plotted against each other

By examining this same scatterplot but just for one MSA, the patterns are more evident, and much more contrasted, as when comparing the scatter plots for Delaware Valley, one of the bigger MSAs in Figure 16, to Charleston in Figure 17, one of the smallest MSAs. Examining the scatterplots for these cities individually allows for identification of relationships between variables in their unique geographical context. Comparing the scatterplots between the two MSAs allows for us to see whether patterns observed in one location hold true in the other. Looking at all the information compiled into one place as in Figure 15 provides the context in which the individual MSAs can be positioned, as well as to identify whether patterns noticed for the general database are also applicable to the individual MSA and vice-versa.

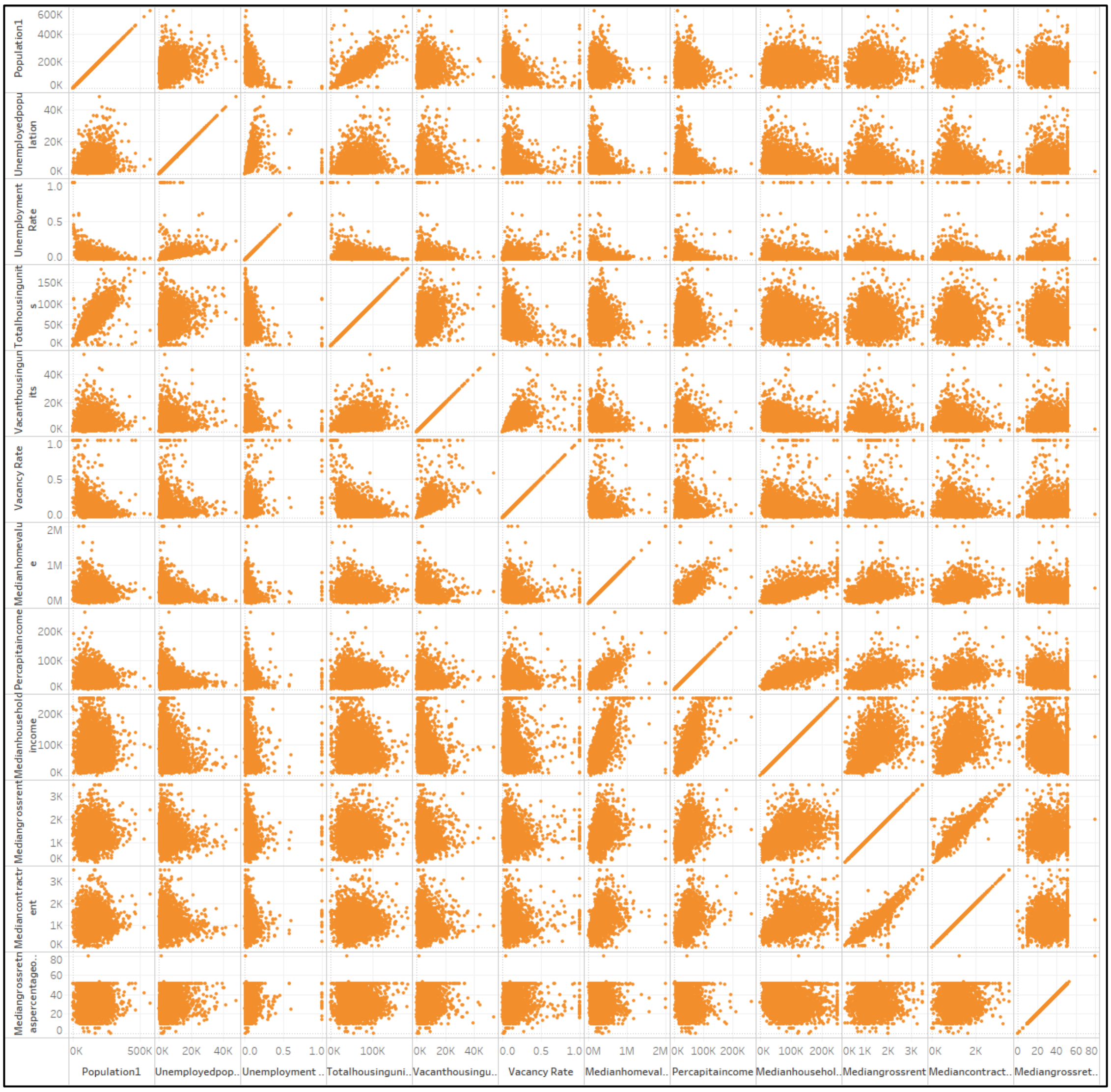


Figure 16: Scatter plots of KPIs for Delaware Valley

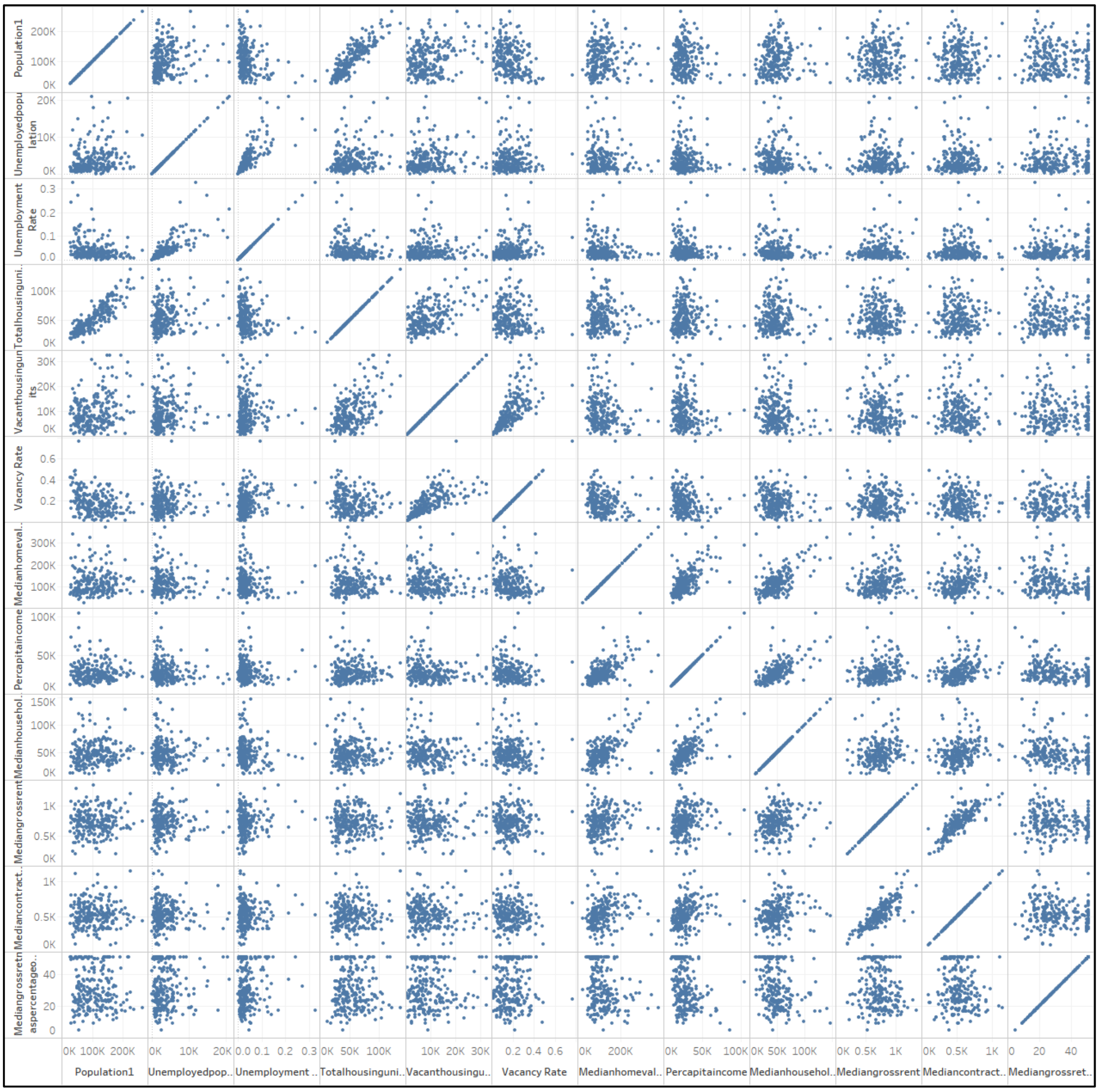

Figure 17: Scatter plots of KPIs for Charleston

Descriptive statistics can also be used to evaluate the distribution of data for individual moments in the database. For example, developing box plots for all of the cities in the database and comparing them with box plots for the individual cities for the most recent year of data, show the viewer that the distribution of the second and third quartile for the median gross rent as a percentage of household income remained relatively stable across all of the individual MSAs in 2020, even while the median household income varies quite a bit in its average and quartile values across different MSAs. This is seen in Figure 18 below.

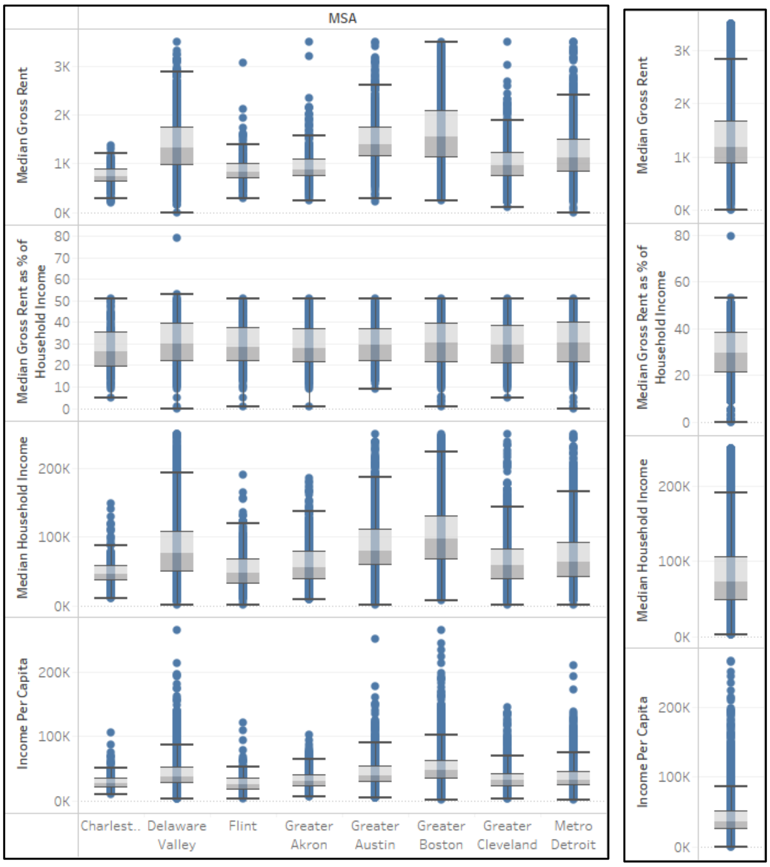


Figure 18: Box-and-whisker plots for rent and income-related indicators

Violin plots are another popular form of descriptive statistics analysis, a bit like a combination of box and whisker plots with histograms. By looking at the distribution of gross rent across the whole database in Figure 19, it looks like the gross rent was largely gathered around USD 1000 monthly. However, Figure 20 shows that the distribution of gross rent varies quite widely across individual MSAs, resulting in some areas with gross rent concentrated around certain values and others which are spread out along a much greater range, potentially indicating greater inequality.

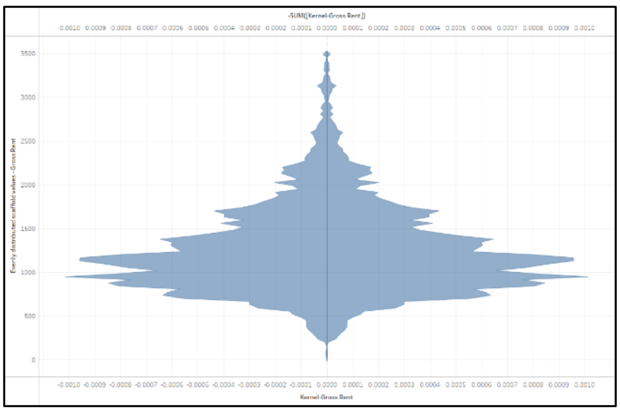

Figure 19: Violin plot of gross rent for Block Groups in all MSAs in 2020

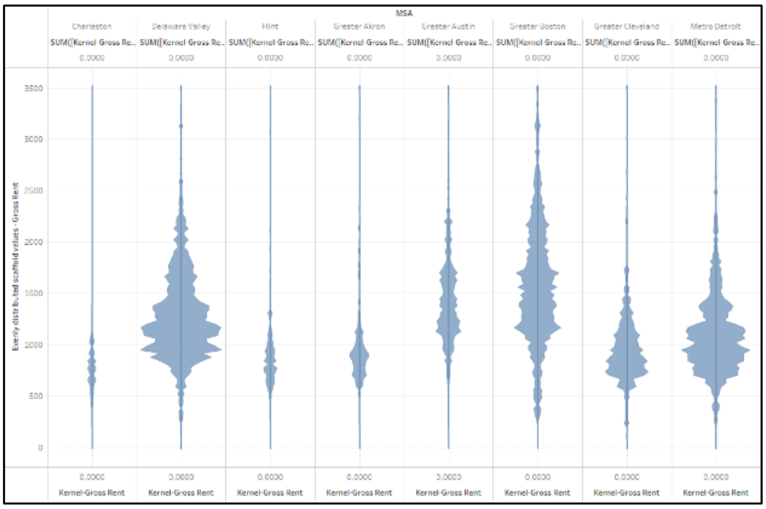


Figure 20: Contrasting violin plots of gross rent for Block Groups in individual MSAs in 2020

In general, descriptive statistics are helpful for providing overviews of the dataset. They can be used to describe the distribution of data for individual as well as agglomerated units of measure, and also serve as first forages into the identification of trends and patterns to be quantified.

## 5.3 Time Series Trends

When examining the data over the time series, it is interesting to look for trends over time, and to see whether there are global or local trends, or both in the data. Even without extracting models for forecasting values, it is interesting to see if the trends that are observed globally are also applicable in local and smaller aggregations, and vice versa. This can help identify areas which may be neglected or other separate from the global average if it is evident that the time series of the location for an indicator follows a different trajectory from the rest.

### 5.3.1 Global Trends

In Figure 21, it looks like globally, all the indicators were mostly increasing in trend, except in the very last year. It is validating to see indicators of similar category following similar trendlines, and interesting to see where there are or are not similar trends in other categories; for example, population and civilian labor force follow similar trend lines globally, which makes sense. However, although contract rent and gross rent as a percentage of household income follow similar trajectories, even in 2011 where both lines experience a dip, gross rent does not follow their patterns, creating interest in exploring the reason for this discrepancy.

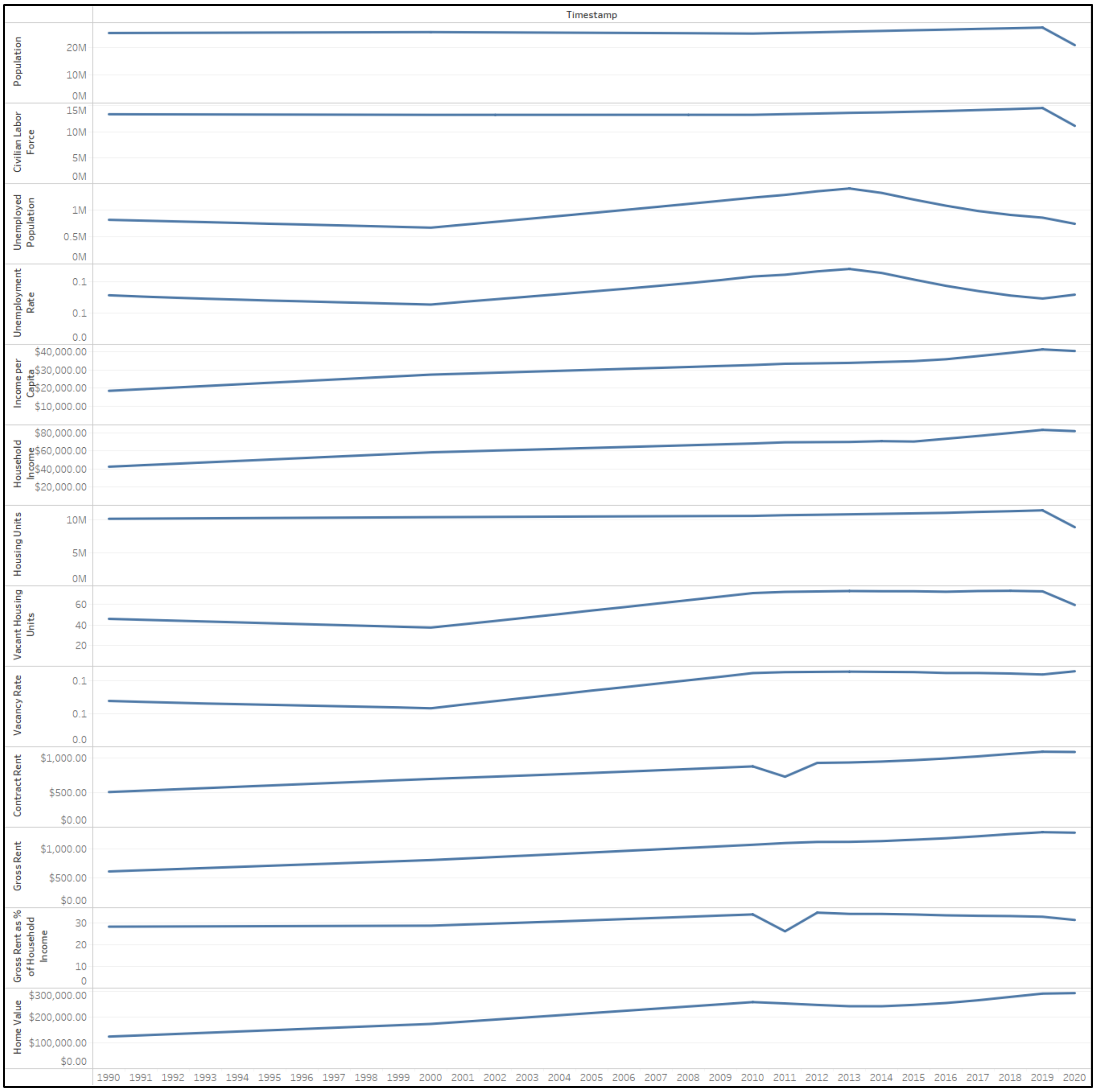


Figure 21: Global time series for Census KPIs

### 5.3.2 Differences across MSAs

After observing the global trends in the indicators over the time series, it is interesting to examine the trajectory of indicator development at a difference urban scale. In the case of Figure 22, the trajectories for each MSA can be seen for each KPI. Although it looks like most of the curves are similar for each indicator, it is interesting to note where for some MSAs the curves are more defined. For example, in unemployed population, there is a sharp increase in numbers in Delaware Valley in 2014, while in Greater Austin, the increase was much milder.

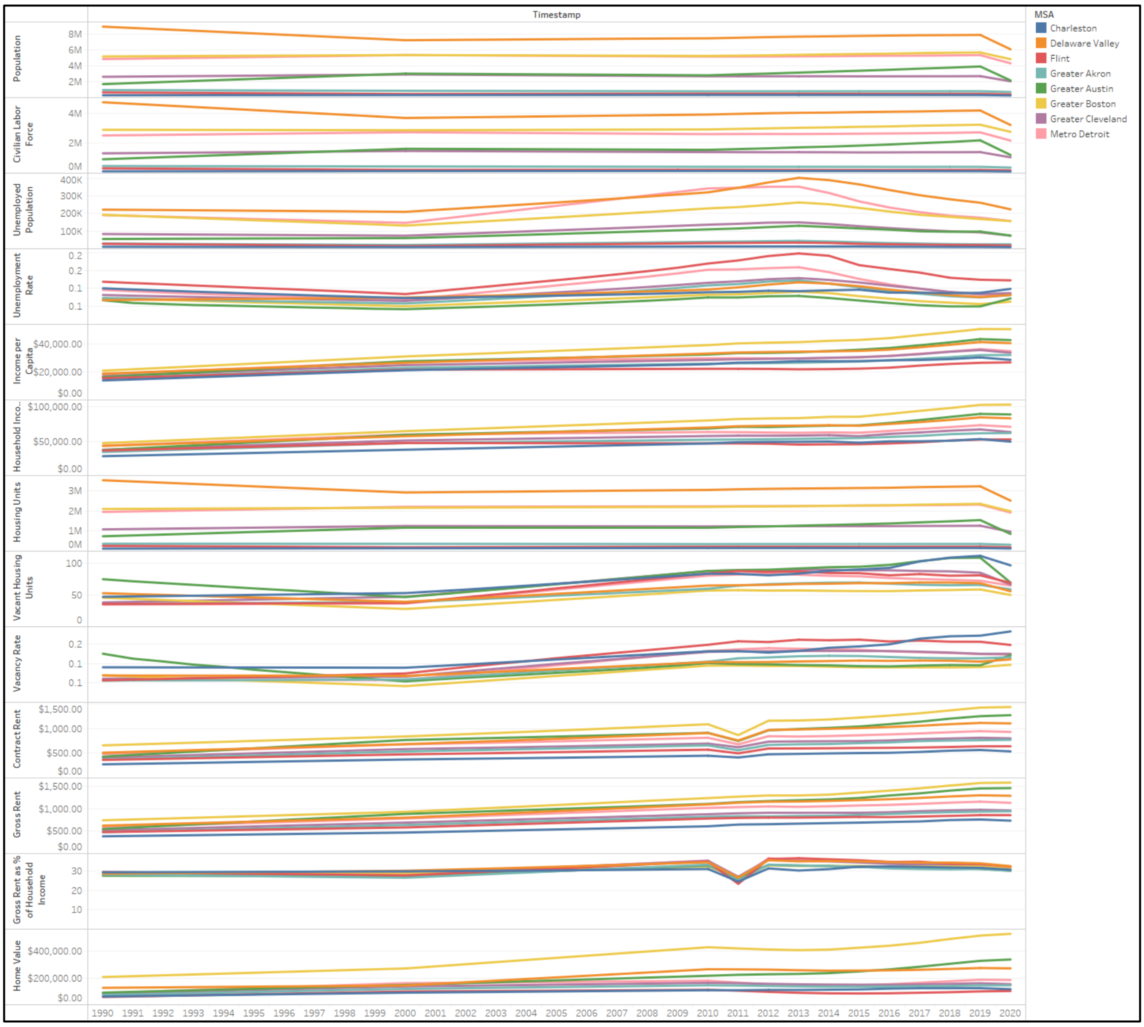


Figure 22: Time series for Census KPIs for individual MSAs

This kind of analysis at smaller aggregations is very interesting and useful for distinguishing traits and circumstances unique to specific locations; the same exercise above could be repeated at the Zip Code, Census Tract, and Block Group level. However, due to the larger quantities of the smaller aggregations of data, it might be harder to distinguish between the individual locations. In this case, it could be more useful to map the values using GIS to better visualize and distinguish major changes in areas from over the time series.

### 5.3.3 Time Series Causal Mechanisms and Delays

Another interesting analysis to be performed when collecting data in time series and at different urban scales is the examination of the temporal relationship between indicators, both on global and local scales. For example, when examining the changes in median household income and gross rent, it is seen in Figure 23 that across all MSAs in the database, household income and gross rent tend to be positively correlated, as they grow together over the time series. On the other hand, when examining the year-on-year percentage changes, the rate of change is different for each indicator. It is interesting to note that not only is household income more volatile in the percentage of change experienced from year to year, but also it appears that the changes in slope in gross rent seem to lag slightly behind those of household income. That is, when household income changes a lot, gross rent changes also, but tends to follow the rate of change a year or two after the change in income.

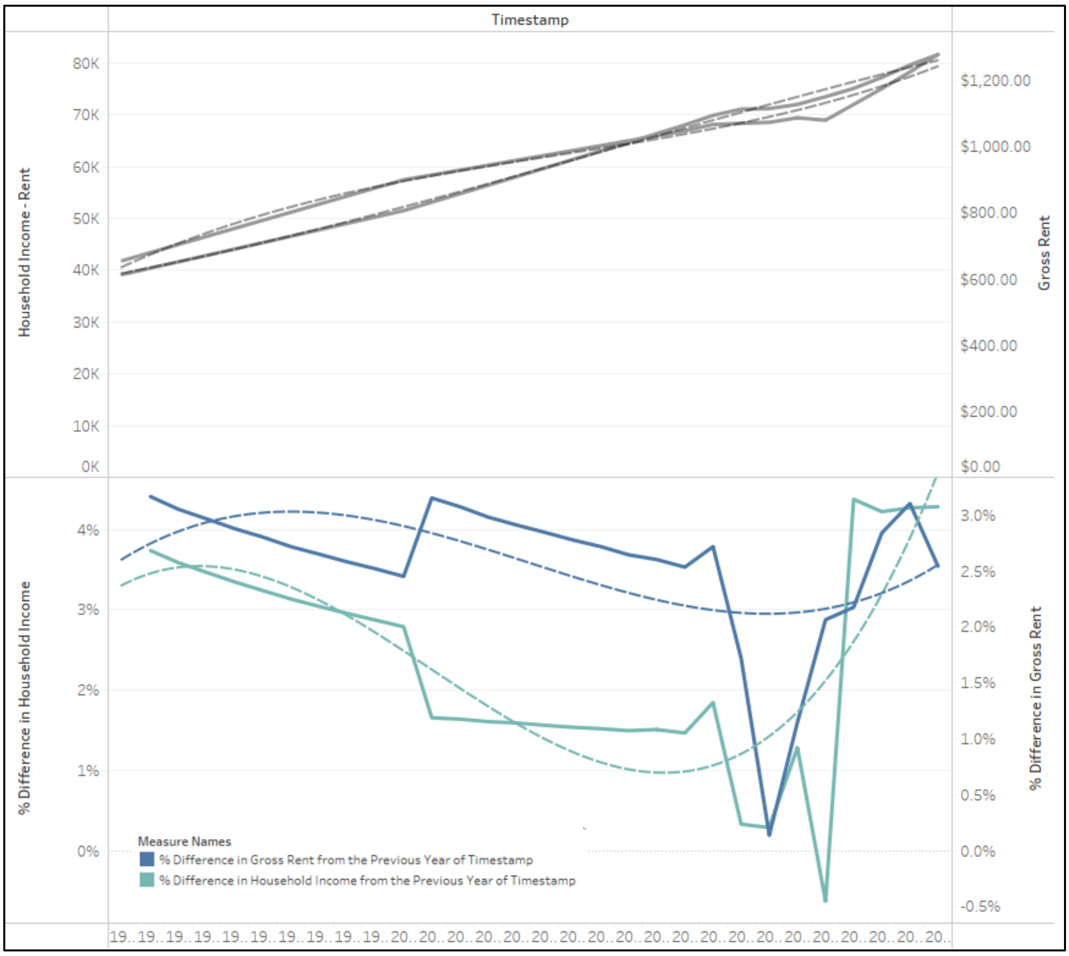


Figure 23: Annual global changes in household income and gross rent

Examining the same indicators at the scale of MSAs, similar trends are observed in Greater Boston as well as in Greater Austin. However, it can be seen that the fluctuations in year-on-year changes in household income and gross rent in Boston are closer together in scale compared to the global average. Here in Figure 24, it appears that changes in the slope of household income actually lag changes in gross rent by a bit.

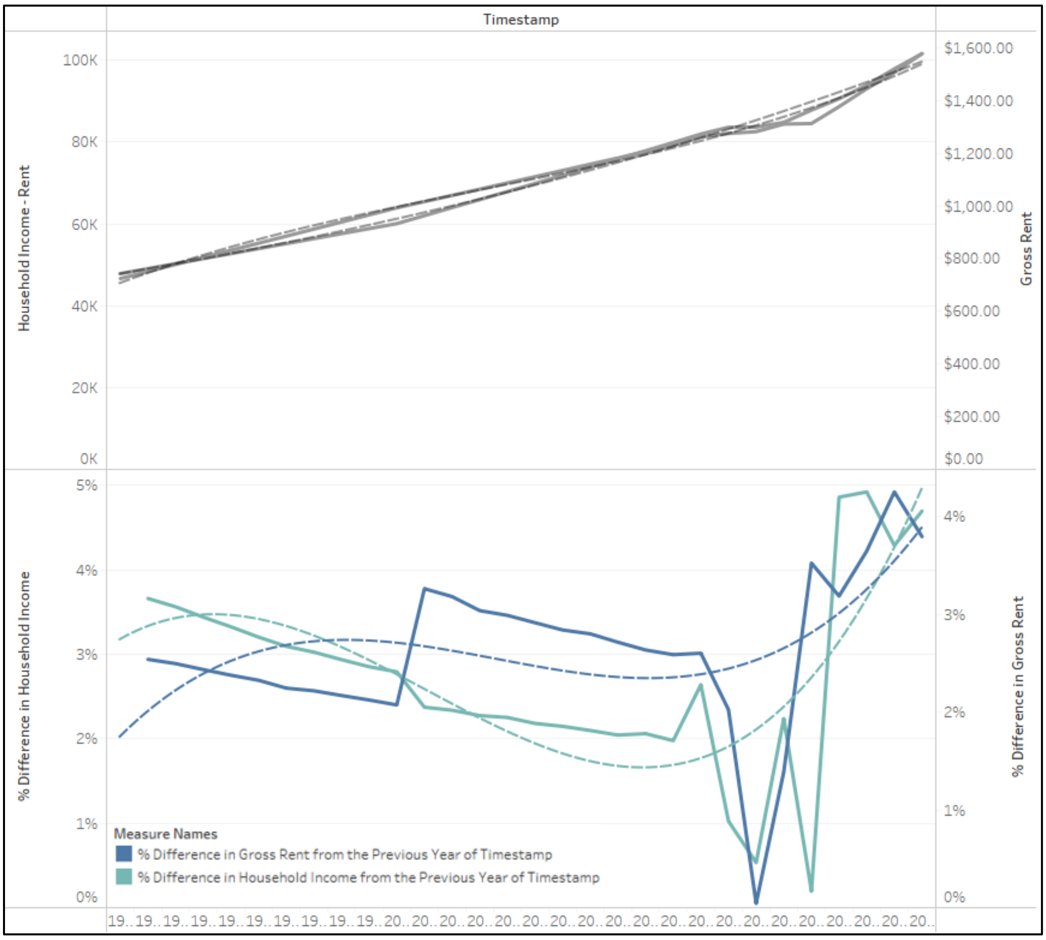


Figure 24: Annual changes in household income and gross rent in Greater Boston

Greater Austin, on the other hand, displays a trend more like the global average, although the curves are slightly different (this may be indicative of a different time cycle of urban change for this city). Figure 25 shows that although the household income and gross rent are not are linearly correlated as compared to the global average or even Greater Boston, there is still a positive correlation between the two factors. When examining the change in slope, it appears that there is a very faint, if any lag in change in gross rent from the previous year compared to changes in household income. This may be indicative of an economy in which median rent and household income are more closely linked together and less responsive to external pressures such as growing population or home values (although these would need to be examined together to verify this).

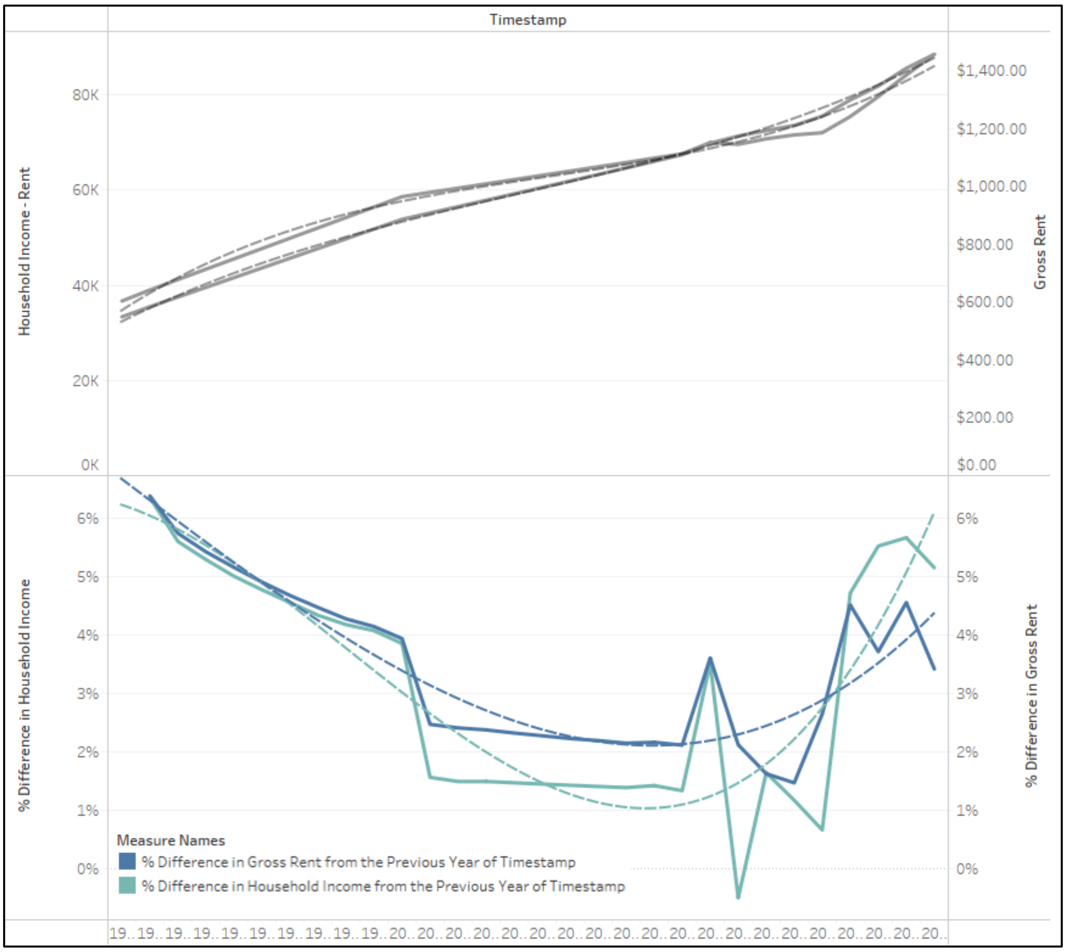


Figure 25: Annual changes in household income and gross rent in Greater Austin

Like the other time series trends, this analysis can also be performed at smaller scales of urban aggregation (county, tract, zip code, block group) to confirm whether these patterns hold across the larger geography, or if there are certain areas which more strongly influence others in the larger agglomerations. Having a database that allows for this flexibility in scale of assessment would thus be advantageous for pinpointing more specifically those areas exhibiting stronger behaviors than others.

## 5.4 Cross-Sectional Relationships

In addition to evaluating the evolution of urban performance over time, it is also valuable to assess cross-sections of data – first, to obtain reference points of comparison, and second to be able to make comparisons to other indices or reports which may have been produced for specific years of analysis. When the database is examined at a cross-section in time, it provides an opportunity to search for innate relationships between indicators. While in time series, the change in one indicator may effect changes in another, in cross-sections, the data can effectively be examined as a “still shot” in which only the direct relationships between factors in that slice of time are analyzed against each other.

### 5.4.1 Relationships between Variables

In cross-sectional analyses, it is interesting to examine closely the relationship between variables. For example, it is interesting to explore the log-log relationship between population and the other census KPIs. Given that Bettencourt et al. showed the power scaling relationship between population and urban output [32], it is interesting to see whether there is a relationship between population size and the other urban indicators included in the database. Positive correlation evident with labor force, unemployed population, and housing units, and negative correlations with unemployment rate and vacancy rate. There is also a slight positive correlation between the log of population and the log of median household income.

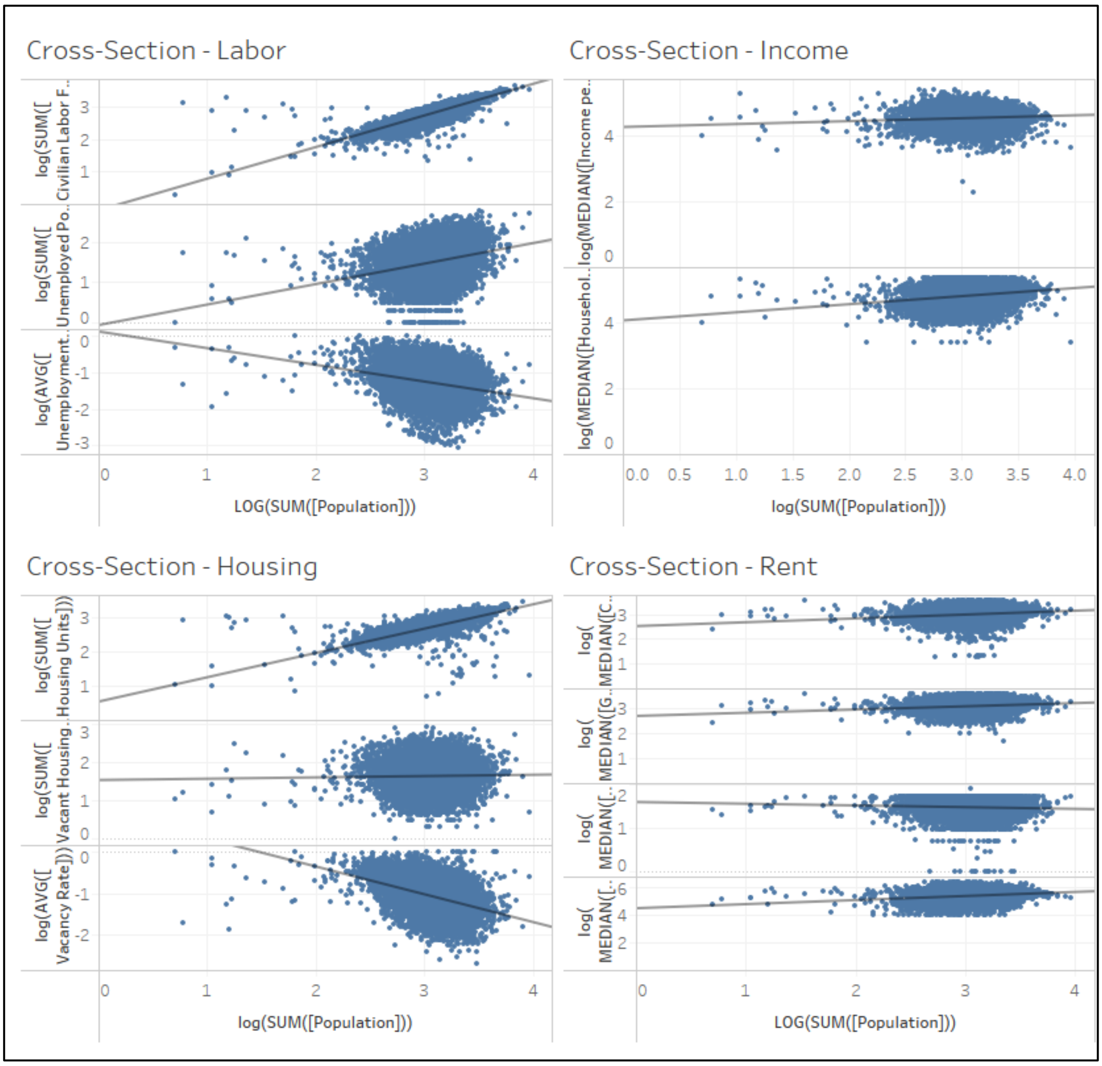


Figure 26: Global relationship between population and other KPIs in 2020

### 5.4.2 Differences Across MSAs

Another type of analysis to explore is the relationship between population and livability index category scores. Interestingly, it looks like there are shapes of correlation associated with population and livability categories of composite score, engagement, environment, health, neighborhood, opportunity, and transportation (from the top left down in Figure 27).


Figure 27: Log-log relationship between 2015 population and livability scores for MSAs

When examining the KPIs from the Census in comparison with the different categories of the Livability Index, Figure 28 and Figure 29 show that there are indeed some correlations found between indicators and some score categories. The former shows that there appears to be a few non-insignificant correlations between KPI and livability score (by category more so than the composite score). The r-values and p-values for the calculated trend lines can be further examined to determine how close the correlations are. One observation to note in this case is that some of the livability index score categories include some of the census indicators in the constructed database; therefore, it should be expected for some categories to show stronger degrees of correlation.

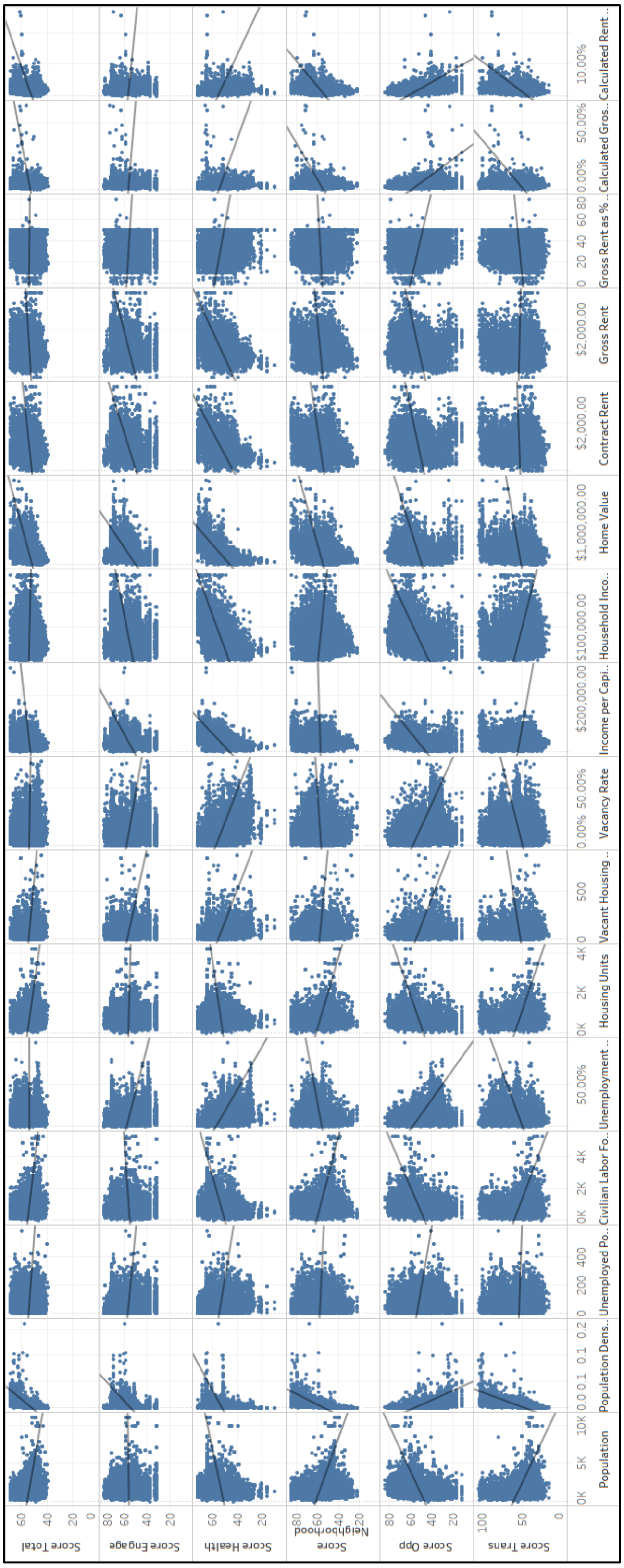


Figure 28: Livability index scores versus indicators at the Block Group level for all MSAs in the database

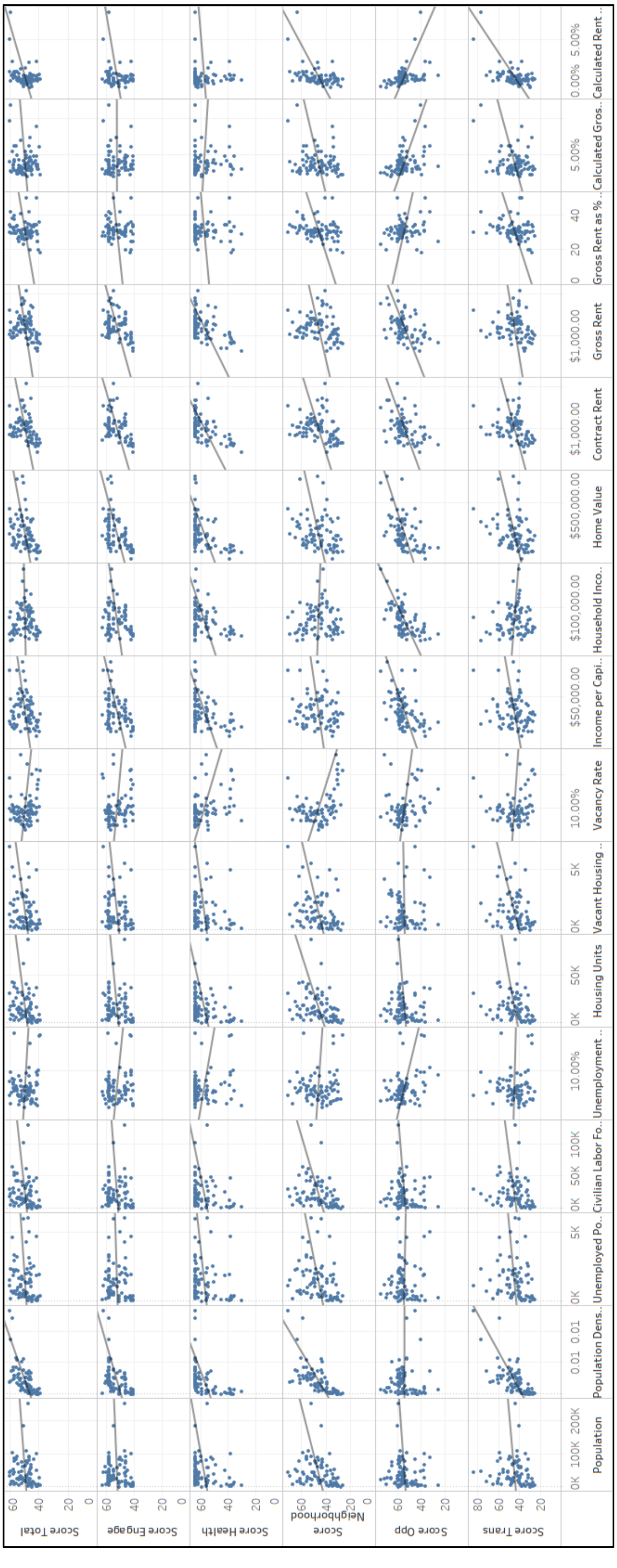

Figure 29: Livability index scores versus indicators at the zip code level for Greater Austin

### 5.4.3 Cluster Analyses for Changes Over Time

Last but not least, having a consistent database with spatiotemporal indicators provides the opportunities to measure indicators and their relationships to their surrounding spaces over time. The use of ArcGIS Pro's cluster and outlier analysis upon changes in livability scores between 2015 and 2022 can be used to identify cluster types of change, which can help point out if there are any areas that could possibly have intentional impacts on surrounding areas, and to identify locations to keep an eye on. Figure 30 shows the results of a cluster and outlier analysis performed upon the changes in livability scores experienced in the neighborhoods of Greater Austin between 2015 and 2022. It is evident that the surrounding area of the MSA improved, while there were mixed results in the city center.

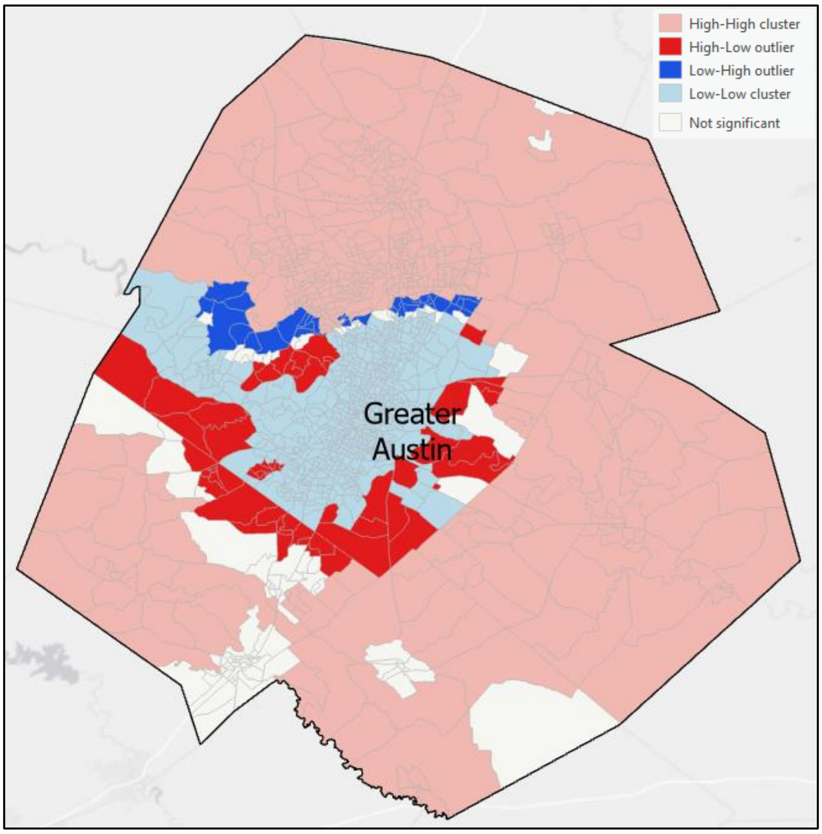


Figure 30: Clusters and outliers of changes in livability scores between '15-'22 in Austin

Each one of these types of analyses is not interesting just in and of themselves, but because they are able to be performed upon the same database of information. As a result, the findings from each descriptive statistics, time series trends, and cross-sectional relationships can be evaluated not just within themselves but also in the context of the other findings. The results discussed above are just a demonstration of the way that developing a database and a spatiotemporal framework can enrich the urban evaluation process for smart city planners and enhance their assessment of opportunities for regeneration.

# 6 Conclusion & Future Work

This final chapter concludes the work by summarizing the results and discussing some potential practical applications. In addition, comments on the limitations of the study are included, along with some proposed recommendations for further development of this work.

## 6.1 Conclusion

This thesis contributed to the development of a conceptual framework of smart city planning through urban performance evaluation through a spatiotemporal lens in the context of decay and regeneration. By diving into the existing literature and attempting to create a consistent database using a detailed yet simplified methodology for multi-sourcing and imputing missing data, a deeper understanding and appreciation of the complexities of urban data availability and analyses was obtained.

A thorough methodology was presented for cobbling together a consistent database even with limited and disparate resources. This reviewed the rationale for simplifying and streamlining the data formatting and imputation process while ensuring that the information remains comparable even across changing urban boundaries. While validation of this methodology by comparing its accuracy to ground truth data would be helpful to authenticating the steps designed for this process, this thesis work illustrates the first steps of a venture into creating a systematic procedure for simplifying urban analysis while maintaining the comprehensiveness and encompassing the complexity in scope. The methodology developed and carried out for building the database can be relatively easily replicated to expand the scope of the database to other MSAs as well as to include other census indicators into the analysis. Many of the considerations in the data engineering process have already been dealt with in the work of the thesis (and documented), hence facilitating future endeavors.

From the experiences of the extensive literature review, creation of an urban performance database, and the methodological framework that was established for evaluating spatiotemporal data on different dimensional scales, it is evident that it is very challenging to plan a city smartly in a way that encompasses all of the different influences that different factors have upon the performance, or livability, or attractiveness of a particular location, while also considering not only the physical and economic impacts, but also the social replications of these efforts. Although very preliminary, the exploration of the created database revealed some interesting patterns and new perspective for exploring urban indicator data over time, space, and in relation with other measures of urban performance such as the AARP livability index.

In conclusion, this thesis presented the case for adopting a spatiotemporal assessment framework for evaluating urban indicators over multiple scales of urban aggregation and over time in order to provide a tool for comprehensive urban performance assessment. In the resulting database that was created through this process, the initial data exploration resulted in some interesting observations made, inviting opportunities to further evaluate relationships between indicators and in contexts which may not have previously been considered, even with a limited or seemingly simplistic selection of indicators.

## 6.2 Practical Applications

Although the creation of a database of so much detail is normally an arduous undertaking, the methodology and simplifications presented in this work may facilitate the process for future evaluations of other locations and with other common indicators. The theoretical and practical concepts reviewed have shown the value of using a spatiotemporal framework for assessing urban conditions to optimize decision-making in urban governance. Compiling and analyzing data in these different spatial aggregations and time series may help stakeholders identify not only existing opportunities for growth and regeneration but also to spot patterns in performance metrics over time which can be used for better planning. It is also possible that this kind of analysis will help analysts understand urban dynamics in a more visual yet still quantifiable manner that will help them model the changes occurring within their city and ultimately help look out for upcoming changes (for example, response of the real estate market and rental prices to economic growth).

## 6.3 Limitations & Challenges

The following section outlines some of the main limitations and challenges involved during the process of this thesis.

In general, while data-driven analyses are the backbone of decision, it is important to keep in mind the complexities and challenges behind constructing data, and the different ways in which data can be represented. Even while data is used to emphasize objectivity, accuracy, and neutrality [16], the process of constructing the database made evident just how manual the process can be, and that the decisions made with respect to how to treat and clean the data can also impact the outcome of the analyses.

### 6.3.1 Crosswalks

Crosswalks were only available from 1990-2010 and 2010-2020 at block level or at tract level. Because they had to be aggregated into Block Groups, the methodology selected was to use the maximum value of a block within a Block Group as the method of aggregation. However, it was later noticed that there were large differences between the 1990-2019 data and the 2020 data, which although initially (and could be partially) explained by the impacts of the COVID-19 pandemic which affected not only the indicator values themselves but also the collection methodology due to the complications of collecting census information from households during that time. Since the 2020 data did not require any crosswalking, there was no aggregation step involved. It is possible that there may have been an error in the simplified crosswalking methodology used for the previous years which could impact the comparison with the last year of the dataset.

### 6.3.2 Data Scarcity

The availability of data across time, geography, and in the desired granularity presented many challenges to this research and will continue to be a challenge for future endeavors. As many urban analysts have pointed out, the availability and the consistency of data presents a challenge to continuing this research and applying it to other locations.

A simplified approach works given the limitations in time for this research as well as in order to commit to the “Timely” part of a SMART approach to assessing urban performance. However, it must be understood that with additional simplifications increases the risk of decreasing the accuracy of the data and resulting in incorrect correlations found or conclusions drawn. Furthermore, with each step taken to account for missing data potentially results in a further departure from the “true” data, which can further affect the outcomes of the analysis.

### 6.3.3 Variables

It is clear that the evaluation of a city is very complex and requires the integration of many different variables in the evaluation to create indices such as the Livability Index. Although in this thesis we attempted to simplify this process, given that all these indicators are also included in the livability index may cause for concern in the truthfulness of the evaluation. Also, it is not clear whether the selected indicators are sufficient to measure the accuracy of the proxy index, or if even they are the most relevant indicators to consider for a "simplified" assessment, if such is even realistically possible.

## 6.4 Future Work

The bulk of this thesis ended up being centered in the creation of the database. As a result, there are quite a few future recommendations associated with this work. The most pressing recommendation is to more deeply analyze the database developed in this thesis, and the model the KPIs with the Livability Index scores within both local and global spatial contexts.

Another interesting future effort would be to use a subset of the database constructed in this thesis, or to expand the database following the methodology developed, and use this to train an algorithm that could better predict missing values for imputation and/or also to predict urban performance metrics for other locations in other years. Furthermore, since the Livability Index scores were also downloaded for 2022, it would be interesting to a) build a forecast model within the database to predict 2022 values using the database values; or b) add more recently published KPI data (2021 and 2022) to the database and build a model to verify any relationships between the KPIs and the AARP Livability Score.

From the perspective of the data itself, there are several improvements that could be implemented in future works. In particular, the cross-walking methodology should be revised, and a data verification step included to ensure that the data is correctly translated from source years to target years. Data checks should be carefully considered and implemented to assure data integrity during the data wrangling process as well.

Lastly, given that one of the intended contributions of this thesis was to contribute to a methodology for multi-sourcing and estimating data to create a consistent database, future efforts could try to expand the created database or even to construct new databases that also encompass spatiotemporal urban indicators at multiple scales of urban aggregation which could be used to derive insights into new locations both in the U.S. as well as abroad. For example, the MSAs in the United States are similar to Larger Urban Zones in Europe [34]; it would be interesting to compare urban performance across a larger set of geographies and to explore similarities and differences in spatiotemporal trends.

# Bibliography


[1] E. P. Trindade, M. P. F. Hinnig, E. M. da Costa, J. S. Marques, R. C. Bastos, and T. Yigitcanlar, “Sustainable development of smart cities: a systematic review of the literature,” *J. open innov.*, vol. 3, no. 1, p. 11, Dec. 2017, doi: 10.1186/s40852-017-0063-2.

[2] V. Albino, U. Berardi, and R. M. Dangelico, “Smart Cities: Definitions, Dimensions, Performance, and Initiatives,” *Journal of Urban Technology*, vol. 22, no. 1, pp. 3–21, Jan. 2015, doi: 10.1080/10630732.2014.942092.

[3] G. T. Doran, “There’s a S.M.A.R.T. way to write managements’s goals and objectives.,” *Management Review*, vol. 70, no. 11, p. 35, Nov. 1981.

[4] A. Brown and R. F. Kirby, “Measuring Urban Performance,” *Journal of Cybernetics*, vol. 1, no. 4, pp. 32–54, Jan. 1971, doi: 10.1080/01969727108542901.

[5] C. J. L. Balsas, “Measuring the livability of an urban centre: an exploratory study of key performance indicators,” *null*, vol. 19, no. 1, pp. 101–110, Feb. 2004, doi: 10.1080/0269745042000246603.

[6] F. Léautier, *Governance and the City: An Empirical Exploration Into Global Determinants of Urban Performance*. World Bank, World Bank Institute, 2005. [Online]. Available: https://books.google.es/books?id=7sivjpGAZd8C

[7] A. Caragliu and C. Del Bo, “Smartness and European urban performance: assessing the local impacts of smart urban attributes,” *Innovation: The European Journal of Social Science Research*, vol. 25, no. 2, pp. 97–113, Jun. 2012, doi: 10.1080/13511610.2012.660323.

[8] J. Mavrič and V. Bobek, “Measuring Urban Development and City Performance,” in *Perspectives on Business and Management*, V. Bobek, Ed. InTech, 2015. doi: 10.5772/61063.

[9] J. M. Leach, P. A. Braithwaite, S. E. Lee, C. J. Bouch, D. V. L. Hunt, and C. D. F. Rogers, “Measuring urban sustainability and liveability performance: the City Analysis Methodology,” *IJCAST*, vol. 1, no. 1, p. 86, 2016, doi: 10.1504/IJCAST.2016.081296.

[10] S. Rufat, “Spectroscopy of Urban Vulnerability,” *Annals of the Association of American Geographers*, vol. 103, no. 3, pp. 505–525, May 2013, doi: 10.1080/00045608.2012.702485.

[11] Á. Corredor-Ochoa, C. Antuña-Rozado, J. Fariña-Tojo, and J. Rajaniemi, “Challenges in assessing urban sustainability,” in *Urban Ecology*, Elsevier, 2020, pp. 355–374. doi: 10.1016/B978-0-12-820730-7.00019-7.

[12] United Nations Department of Economic and Social Affairs/Population Division, “World Urbanization Prospects: The 2018 Revision,” United Nations, New York, ST/ESA/SER.A/420, 2019. Accessed: May 15, 2022. [Online]. Available: https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf

[13] Google, “smart city,” *Google Trends*. https://trends.google.com/trends/explore?date=all&q=smart%20city (accessed Jun. 15, 2022).

[14] S. Allwinkle and P. Cruickshank, “Creating Smart-er Cities: An Overview,” *Journal of Urban Technology*, vol. 18, no. 2, pp. 1–16, Apr. 2011, doi: 10.1080/10630732.2011.601103.
[15] M. Batty *et al.*, “Smart cities of the future,” *Eur. Phys. J. Spec. Top.*, vol. 214, no. 1, pp. 481–518, Nov. 2012, doi: 10.1140/epjst/e2012-01703-3.
[16] T. Shelton, M. Zook, and A. Wiig, “The ‘actually existing smart city,’” *CAMRES*, vol. 8, no. 1, pp. 13–25, Mar. 2015, doi: 10.1093/cjres/rsu026.
[17] H. Chourabi *et al.*, “Understanding Smart Cities: An Integrative Framework,” in *2012 45th Hawaii International Conference on System Sciences*, Maui, HI, USA, Jan. 2012, pp. 2289–2297. doi: 10.1109/HICSS.2012.615.
[18] R. G. Hollands, “Will the real smart city please stand up?: Intelligent, progressive or entrepreneurial?,” *City*, vol. 12, no. 3, pp. 303–320, Dec. 2008, doi: 10.1080/13604810802479126.
[19] N. Komninos, “Intelligent cities: Variable geometries of spatial intelligence,” *Intelligent Buildings International*, vol. 3, no. 3, pp. 172–188, Jul. 2011, doi: 10.1080/17508975.2011.579339.
[20] M. Angelidou, “Smart city policies: A spatial approach,” *Cities*, vol. 41, pp. S3–S11, Jul. 2014, doi: 10.1016/j.cities.2014.06.007.
[21] G. C. Lazaroiu and M. Roscia, “Definition methodology for the smart cities model,” *Energy*, vol. 47, no. 1, pp. 326–332, Nov. 2012, doi: 10.1016/j.energy.2012.09.028.
[22] S. Koutra, V. Becue, and C. S. Ioakimidis, “Searching for the ‘smart’ definition through its spatial approach,” *Energy*, vol. 169, pp. 924–936, Feb. 2019, doi: 10.1016/j.energy.2018.12.019.
[23] Z. Allam and P. Newman, “Redefining the Smart City: Culture, Metabolism and Governance,” *Smart Cities*, vol. 1, no. 1, pp. 4–25, Jul. 2018, doi: 10.3390/smartcities1010002.
[24] R. Giffinger, C. Fertner, H. Kramar, R. Kalasek, N. Pichler-Milanović, and E. Meijers, “Smart cities: Ranking of European medium-sized cities,” Centre of Regional Science (SRF), Vienna University of Technology, Research Project, Oct. 2007. Accessed: May 25, 2021. [Online]. Available: http://www.smart-cities.eu/download/smart_cities_final_report.pdf
[25] A. Caragliu, C. Del Bo, and P. Nijkamp, “Smart Cities in Europe,” *Journal of Urban Technology*, vol. 18, no. 2, pp. 65–82, Apr. 2011, doi: 10.1080/10630732.2011.601117.
[26] M. Angelidou, “Smart cities: A conjuncture of four forces,” *Cities*, vol. 47, pp. 95–106, Sep. 2015, doi: 10.1016/j.cities.2015.05.004.
[27] M. Alberti, “Measuring urban sustainability,” *Environmental Impact Assessment Review*, vol. 16, no. 4–6, pp. 381–424, Jul. 1996, doi: 10.1016/S0195-9255(96)00083-2.
[28] A. Repetti and G. Desthieux, “A Relational Indicatorset Model for urban land-use planning and management: Methodological approach and application in two case studies,” *Landscape and Urban Planning*, vol. 77, no. 1–2, pp. 196–215, Jun. 2006, doi: 10.1016/j.landurbplan.2005.02.006.
[29] S. Zygiaris, “Smart City Reference Model: Assisting Planners to Conceptualize the Building of Smart City Innovation Ecosystems,” *J Knowl Econ*, vol. 4, no. 2, pp. 217–231, Jun. 2013, doi: 10.1007/s13132-012-0089-4.
[30] J. Hui and S. Lim, “An Analytic Hierarchy Process (AHP) Approach for Sustainable Assessment of Economy-Based and Community-Based Urban Regeneration: The Case of South Korea,” *Sustainability*, vol. 10, no. 12, p. 4456, Nov. 2018, doi: 10.3390/su10124456.

[31] S. Marshall, Y. Gong, and N. Green, "Urban compactness: indicators of a property distinct from density," 2010. Accessed: Apr. 04, 2022. [Online]. Available: https://livrepository.liverpool.ac.uk/3001755/1/Marshall%20Gong%20Green%202010%20AESOP%20Compactness%20%5BMS%5D.pdf

[32] L. M. A. Bettencourt, J. Lobo, D. Strumsky, and G. B. West, "Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities," *PLoS ONE*, vol. 5, no. 11, p. e13541, Nov. 2010, doi: 10.1371/journal.pone.0013541.

[33] L. G. A. Alves, R. S. Mendes, E. K. Lenzi, and H. V. Ribeiro, "Scale-Adjusted Metrics for Predicting the Evolution of Urban Indicators and Quantifying the Performance of Cities," *PLoS ONE*, vol. 10, no. 9, p. e0134862, Sep. 2015, doi: 10.1371/journal.pone.0134862.

[34] E. Arcaute, E. Hatna, P. Ferguson, H. Youn, A. Johansson, and M. Batty, "Constructing cities, deconstructing scaling laws," *J. R. Soc. Interface.*, vol. 12, no. 102, p. 20140745, Jan. 2015, doi: 10.1098/rsif.2014.0745.

[35] V. C. Yang, A. V. Papachristos, and D. M. Abrams, "Modeling the origin of urban-output scaling laws," *Phys. Rev. E*, vol. 100, no. 3, p. 032306, Sep. 2019, doi: 10.1103/PhysRevE.100.032306.

[36] L. Bettencourt and G. West, "A unified theory of urban living," *Nature*, vol. 467, no. 7318, pp. 912–913, Oct. 2010, doi: 10.1038/467912a.

[37] R. C. Weaver and S. Bagchi-Sen, "Spatial analysis of urban decline: The geography of blight," *Applied Geography*, vol. 40, pp. 61–70, Jun. 2013, doi: 10.1016/j.apgeog.2013.01.011.

[38] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann, "The Product Space Conditions the Development of Nations," *Science*, vol. 317, no. 5837, pp. 482–487, Jul. 2007, doi: 10.1126/science.1144581.

[39] J. Choi, J. C. Seong, B. Kim, and E. L. Usery, "Innovations in Individual Feature History Management—The Significance of Feature-based Temporal Model," *Geoinformatica*, vol. 12, no. 1, pp. 1–20, Mar. 2008, doi: 10.1007/s10707-007-0019-y.

[40] A. M. Dewan and R. J. Corner, "Spatiotemporal Analysis of Urban Growth, Sprawl and Structure," in *Dhaka Megacity*, A. Dewan and R. Corner, Eds. Dordrecht: Springer Netherlands, 2014, pp. 99–121. doi: 10.1007/978-94-007-6735-5_6.

[41] H. Doraiswamy, J. Freire, M. Lage, F. Miranda, and C. Silva, "Spatio-Temporal Urban Data Analysis: A Visual Analytics Perspective," *IEEE Comput. Grap. Appl.*, vol. 38, no. 5, pp. 26–35, Sep. 2018, doi: 10.1109/MCG.2018.053491728.

[42] D. J. Peuquet, "It's about Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems," *Annals of the Association of American Geographers*, vol. 84, no. 3, pp. 441–461, 1994.

[43] W. Zheng, G. Q. Shen, H. Wang, J. Hong, and Z. Li, "Decision support for sustainable urban renewal: A multi-scale model," *Land Use Policy*, vol. 69, pp. 361–371, Dec. 2017, doi: 10.1016/j.landusepol.2017.09.019.

[44] S. S. Rosenthal, "Old homes, externalities, and poor neighborhoods. A model of urban decline and renewal," *Journal of Urban Economics*, vol. 63, no. 3, pp. 816–840, May 2008, doi: 10.1016/j.jue.2007.06.003.

[45] B. T. White, S. M. Sepe, and S. Masconale, "URBAN DECAY, AUSTERITY, AND THE RULE OF LAW+," *Emory Law Journal*, vol. 61, no. 1, 2014, Accessed: Jun. 15, 2022. [Online]. Available: advance-lexis-com.ezproxy1.hw.ac.uk/api/document?collection=analytical-

materials&id=urn:contentItem:5DCW-5K70-00CV-M0V3-00000-00&context=1516831

[46] G. E. Breger, “The Concept and Causes of Urban Blight,” *Land Economics*, vol. 43, no. 4, pp. 369–376, 1967, doi: 10.2307/3145542.

[47] A. Cuthbert, “Urban decay and regeneration: context and issues,” *Journal of Urban Design*, vol. 22, no. 2, pp. 140–143, Mar. 2017, doi: 10.1080/13574809.2017.1288873.

[48] W. Schenkel, “Regeneration Strategies in Shrinking Urban Neighbourhoods—Dimensions of Interventions in Theory and Practice,” *European Planning Studies*, vol. 23, no. 1, pp. 69–86, Jan. 2015, doi: 10.1080/09654313.2013.820089.

[49] M. Degen and M. García, “The Transformation of the ‘Barcelona Model’: An Analysis of Culture, Urban Regeneration and Governance: The cultural transformation of the ‘Barcelona model,’” *International Journal of Urban and Regional Research*, vol. 36, no. 5, pp. 1022–1038, Sep. 2012, doi: 10.1111/j.1468-2427.2012.01152.x.

[50] H. S. Andersen, *Urban Sores: On the interaction between segregation, urban decay and deprived neighbourhoods*, 1st ed. Routledge, 2019. doi: 10.4324/9781315191980.

[51] H. W. Zheng, G. Q. Shen, and H. Wang, “A review of recent studies on sustainable urban renewal,” *Habitat International*, vol. 41, pp. 272–279, Jan. 2014, doi: 10.1016/j.habitatint.2013.08.006.

[52] R. Bolici and L. Mora, “Urban regeneration in the digital era: how to develop Smart City strategies in large european cities,” *TECHNE - Journal of Technology for Architecture and Environment*, pp. 110-119 Pages, Nov. 2015, doi: 10.13128/TECHNE-17507.

[53] Y. Mao, H. Li, and Q. Xu, “The Mode of Urban Renewal Base on the Smart City Theory under the Background of New Urbanization,” *Front. Eng*, vol. 2, no. 3, p. 261, 2015, doi: 10.15302/J-FEM-2015035.

[54] M. Omidipoor, M. Jelokhani-Niaraki, A. Moeinmehr, A. Sadeghi-Niaraki, and S.-M. Choi, “A GIS-based decision support system for facilitating participatory urban renewal process,” *Land Use Policy*, vol. 88, p. 104150, Nov. 2019, doi: 10.1016/j.landusepol.2019.104150.

[55] P. Martí, C. García-Mayor, and L. Serrano-Estrada, “Identifying opportunity places for urban regeneration through LBSNs,” *Cities*, vol. 90, pp. 191–206, Jul. 2019, doi: 10.1016/j.cities.2019.02.001.

[56] L. Hemphill, J. Berry, and S. McGreal, “An Indicator-based Approach to Measuring Sustainable Urban Regeneration Performance: Part 1, Conceptual Foundations and Methodological Framework,” *Urban Studies*, vol. 41, no. 4, pp. 725–755, Apr. 2004, doi: 10.1080/0042098042000194089.

[57] J. Chen, P. Pellegrini, and G. Ma, “Identifying Resettlement Communities’ Urban Regeneration Opportunity Through GIS-based Spatial Analysis in Suzhou Metropolitan Area,” *Urban and Regional Planning*, vol. 6, no. 4, pp. 146–157, Nov. 2021, doi: 10.11648/j.urp.20210604.15.

[58] R. M. Silverman, K. L. Patterson, L. Yin, M. Ranahan, and L. Wu, *Affordable Housing in US Shrinking Cities: From Neighborhoods of Despair to Neighborhoods of Opportunity?* Policy Press, 2016. doi: 10.1332/policypress/9781447327585.001.0001.

[59] L. Hemphill, S. McGreal, and J. Berry, “An Indicator-based Approach to Measuring Sustainable Urban Regeneration Performance: Part 2, Empirical Evaluation and

Case-study Analysis," *Urban Studies*, vol. 41, no. 4, pp. 757–772, Apr. 2004, doi: 10.1080/0042098042000194098.
[60] J.-H. Yu and H.-R. Kwon, "Critical success factors for urban regeneration projects in Korea," *International Journal of Project Management*, vol. 29, no. 7, pp. 889–899, Oct. 2011, doi: 10.1016/j.ijproman.2010.09.001.
[61] T. H. G. Tran, M. Camargo, L. Dupont, and F. Mayer, "A review of methods for modelling shared decisionmaking process in a Smart City Living Lab," in *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC): "Engineering, technology & innovation management beyond 2020: new challenges, new approaches" : conference proceedings*, Institute of Electrical and Electronics Engineers, 2017. Accessed: Aug. 05, 2021. [Online]. Available: http://ieeexplore.ieee.org/servlet/opac?punumber=8269762
[62] N. Khansari, B. G. Silverman, Q. Du, J. B. Waldt, W. W. Braham, and J. M. Lee, "An Agent-Based Decision Tool to Explore Urban Climate & Smart City Possibilities," in *2017 Annual IEEE International Systems Conference (SysCon)*, Piscataway: IEEE, 2017. Accessed: Nov. 14, 2021. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?punumber=7932286
[63] M. Bottero, C. D'Alpaos, and A. Oppio, "Multicriteria Evaluation of Urban Regeneration Processes: An Application of PROMETHEE Method in Northern Italy," *Advances in Operations Research*, vol. 2018, pp. 1–12, Nov. 2018, doi: 10.1155/2018/9276075.
[64] Maria Gracia Riera Pérez and E. Rey, "A multi-criteria approach to compare urban renewal scenarios for an existing neighborhood. Case study in Lausanne (Switzerland)," *Building and Environment*, vol. 65, pp. 58–70, Jul. 2013, doi: 10.1016/j.buildenv.2013.03.017.
[65] M. Jelokhani-Niaraki, A. Sadeghi-Niaraki, and S.-M. Choi, "Semantic interoperability of GIS and MCDA tools for environmental assessment and decision making," *Environmental Modelling & Software*, vol. 100, pp. 104–122, Feb. 2018, doi: 10.1016/j.envsoft.2017.11.011.
[66] S. Praharaj, "Area-Based Urban Renewal Approach for Smart Cities Development in India: Challenges of Inclusion and Sustainability," *UP*, vol. 6, no. 4, pp. 202–215, Nov. 2021, doi: 10.17645/up.v6i4.4484.
[67] J. L. Vigdor, "Is urban decay bad? Is urban revitalization bad too?," *Journal of Urban Economics*, vol. 68, no. 3, pp. 277–289, Nov. 2010, doi: 10.1016/j.jue.2010.05.003.
[68] I. Beretta, "The social effects of eco-innovations in Italian smart cities," *Cities*, vol. 72, pp. 115–121, Feb. 2018, doi: 10.1016/j.cities.2017.07.010.
[69] F. Caprotti, "Critical research on eco-cities? A walk through the Sino-Singapore Tianjin Eco-City, China," *Cities*, vol. 36, pp. 10–17, Feb. 2014, doi: 10.1016/j.cities.2013.08.005.
[70] A. Ganti, "Metropolitan Statistical Area (MSA)," *Investopedia*, Oct. 31, 2021. https://www.investopedia.com/terms/m/msa.asp (accessed Jun. 10, 2022).
[71] M. Schläpfer *et al.*, "The scaling of human interactions with city size," *J. R. Soc. Interface.*, vol. 11, no. 98, p. 20130789, Sep. 2014, doi: 10.1098/rsif.2013.0789.
[72] L. M. A. Bettencourt, J. Lobo, and H. Youn, "The hypothesis of urban scaling: formalization, implications and challenges," 2013, doi: 10.48550/ARXIV.1301.5919.
[73] K. Rossiter, "Understanding Geographic Relationships: Counties, Places, Tracts and More," *United States Census Bureau*, Jul. 31, 2014. https://www.census.gov/newsroom/blogs/random-

samplings/2014/07/understanding-geographic-relationships-counties-places-tracts-and-more.html (accessed Mar. 03, 2022).
[74] AARP Public Policy Institute, “Home - AARP Livability Index,” *AARP*, 2022. https://livabilityindex.aarp.org/ (accessed May 10, 2022).
[75] *Python*. Python, 2021. [Online]. Available: https://www.python.org/downloads/release/python-392/
[76] United States Census Bureau, “Developers,” *United States Census Bureau*, 2022. https://www.census.gov/data/developers.html (accessed Apr. 01, 2022).
[77] *IPUMS National Historical Geographic Information System*. Minneapolis, MN: IPUMS, 2021. [Online]. Available: http://doi.org/10.18128/D050.V16.0
[78] United States Census Bureau, “TIGER/Line Shapefiles,” *United States Census Bureau*, Dec. 16, 2021. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html (accessed Mar. 10, 2022).
[79] Julien Leider, *CensusData*. 2022. [Online]. Available: https://pypi.org/project/CensusData/
[80] United States Census Bureau, “Index of/ (Census.gov),” *United States Census Bureau*, Feb. 03, 2022. https://www2.census.gov/ (accessed Mar. 15, 2022).
[81] *ArcGIS Pro*. ESRI, 2021.
[82] AARP Public Policy Institute, “Zip Code 08062,” *AARP LIVABILITY INDEX*, 2022. https://livabilityindex.aarp.org/search/Mullica%20Hill,%20New%20Jersey%2008062,%20United%20States (accessed Jun. 15, 2022).
[83] “TIGER/Line Shapefiles 2020 Technical Documentation.” U.S. Census Bureau, Feb. 2021. Accessed: Mar. 15, 2022. [Online]. Available: https://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2020/TGRSHP2020_TechDoc.pdf
[84] United States Census Bureau, “Relationship Files,” *United States Census Bureau*, Oct. 28, 2021. https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.html#zcta (accessed Apr. 15, 2022).
[85] IPUMS, “Geographic Crosswalks,” *NATIONAL HISTORICAL GIS*. https://www.nhgis.org/geographic-crosswalks (accessed May 10, 2022).
[86] W. McKinney and others, “Data structures for statistical computing in python,” *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51--56, 2010.
[87] Y. Li and L. E. Parker, “Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks,” *Information fusion*, vol. 15, no. 1, pp. 64–79, 2014.
[88] N. Bokde, M. W. Beck, F. Martínez Álvarez, and K. Kulat, “A novel imputation methodology for time series based on pattern sequence forecasting,” *Pattern recognition letters*, vol. 116, pp. 88–96, 2018.
[89] L. Li and P. Revesz, “Interpolation methods for spatio-temporal geographic data,” *Computers, Environment and Urban Systems*, vol. 28, no. 3, pp. 201–227, May 2004, doi: 10.1016/S0198-9715(03)00018-8.
[90] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.
[91] C. Baweja, “How to Deal with Missing Data in Python,” *Towards Data Science*, Jul. 11, 2020. https://towardsdatascience.com/how-to-deal-with-missing-data-in-python-1f74a9112d93 (accessed May 10, 2022).
[92] S. Kumar, “4 Techniques to Handle Missing values in Time Series Data,” *Towards Data Science*, Apr. 28, 2022. https://towardsdatascience.com/4-techniques-to-handle-missing-values-in-time-series-data-c3568589b5a8 (accessed May 10, 2022).

[93] Bing, “Interpolation methods for time series data,” *D3 View*, Apr. 29, 2022. https://blog.d3view.com/interpolation-methods-for-time-series-data/ (accessed Jun. 12, 2022).
[94] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[95] *Tableau Desktop Professional Edition*. Tableau Software, LLC, 2022. [Online]. Available: http://www.tableau.com
[96] *Tableau Prep Builder*. Tableau Software, LLC, 2022.
[97] ESRI, “How Cluster and Outlier Analysis (Anselin Local Moran’s I) works,” *ArcGIS Pro*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm (accessed Jun. 04, 2022).

# Appendix A

KPIs downloaded at Block Level group with year, survey, reference table, and data source.

| Year | Survey | Table | Indicator | Source |
|---|---|---|---|---|
| **2020** | acs5 | B01003 | Total Population | API |
| **2020** | acs5 | B19049 | Median Household Income in 2020 | API |
| **2020** | acs5 | B19301 | Per Capita Income in 2020 | API |
| **2020** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2020** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2020** | acs5 | B25001 | Total Housing Units | API |
| **2020** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2020** | acs5 | B25058 | Median Contract Rent | API |
| **2020** | acs5 | B25064 | Median gross rent | API |
| **2020** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2020** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2019** | acs5 | B01003 | Total Population | API |
| **2019** | acs5 | B19049 | Median Household Income in 2019 | API |
| **2019** | acs5 | B19301 | Per Capita Income in 2019 | API |
| **2019** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2019** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2019** | acs5 | B25001 | Total Housing Units | API |
| **2019** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2019** | acs5 | B25058 | Median Contract Rent | API |
| **2019** | acs5 | B25064 | Median gross rent | API |
| **2019** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2019** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2018** | acs5 | B01003 | Total Population | API |
| **2018** | acs5 | B19049 | Median Household Income in 2018 | API |
| **2018** | acs5 | B19301 | Per Capita Income in 2018 | API |
| **2018** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2018** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2018** | acs5 | B25001 | Total Housing Units | API |
| **2018** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2018** | acs5 | B25058 | Median Contract Rent | API |

| | | | | |
|---|---|---|---|---|
| **2018** | acs5 | B25064 | Median gross rent | API |
| **2018** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2018** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2017** | acs5 | B01003 | Total Population | API |
| **2017** | acs5 | B19049 | Median Household Income in 2017 | API |
| **2017** | acs5 | B19301 | Per Capita Income in 2017 | API |
| **2017** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2017** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2017** | acs5 | B25001 | Total Housing Units | API |
| **2017** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2017** | acs5 | B25058 | Median Contract Rent | API |
| **2017** | acs5 | B25064 | Median gross rent | API |
| **2017** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2017** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2016** | acs5 | B01003 | Total Population | API |
| **2016** | acs5 | B19049 | Median Household Income in 2016 | API |
| **2016** | acs5 | B19301 | Per Capita Income in 2016 | API |
| **2016** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2016** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2016** | acs5 | B25001 | Total Housing Units | API |
| **2016** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2016** | acs5 | B25058 | Median Contract Rent | API |
| **2016** | acs5 | B25064 | Median gross rent | API |
| **2016** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2016** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2015** | acs5 | B01003 | Total Population | API |
| **2015** | acs5 | B19049 | Median Household Income in 2015 | API |
| **2015** | acs5 | B19301 | Per Capita Income in 2015 | API |
| **2015** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2015** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2015** | acs5 | B25001 | Total Housing Units | API |
| **2015** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2015** | acs5 | B25058 | Median Contract Rent | API |
| **2015** | acs5 | B25064 | Median gross rent | API |
| **2015** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2015** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2014** | acs5 | B01003 | Total Population | API |
| **2014** | acs5 | B19049 | Median Household Income in 2014 | API |

| | | | | |
|---|---|---|---|---|
| **2014** | acs5 | B19301 | Per Capita Income in 2014 | API |
| **2014** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2014** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2014** | acs5 | B25001 | Total Housing Units | API |
| **2014** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2014** | acs5 | B25058 | Median Contract Rent | API |
| **2014** | acs5 | B25064 | Median gross rent | API |
| **2014** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2014** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2013** | acs5 | B01003 | Total Population | API |
| **2013** | acs5 | B19049 | Median Household Income in 2013 | API |
| **2013** | acs5 | B19301 | Per Capita Income in 2013 | API |
| **2013** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | API |
| **2013** | acs5 | B23025 | Unemployed Population 16 Years and Over | API |
| **2013** | acs5 | B25001 | Total Housing Units | API |
| **2013** | acs5 | B25002 | Total Vacant Housing Units | API |
| **2013** | acs5 | B25058 | Median Contract Rent | API |
| **2013** | acs5 | B25064 | Median gross rent | API |
| **2013** | acs5 | B25071 | Median gross rent as a percentage of household income | API |
| **2013** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | API |
| **2012** | acs5 | B01003 | Total Population | GDB |
| **2012** | acs5 | B19049 | Median Household Income in 2012 | GDB |
| **2012** | acs5 | B19301 | Per Capita Income in 2012 | GDB |
| **2012** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | GDB |
| **2012** | acs5 | B23025 | Unemployed Population 16 Years and Over | GDB |
| **2012** | acs5 | B25001 | Total Housing Units | GDB |
| **2012** | acs5 | B25002 | Total Vacant Housing Units | GDB |
| **2012** | acs5 | B25058 | Median Contract Rent | GDB |
| **2012** | acs5 | B25064 | Median gross rent | GDB |
| **2012** | acs5 | B25071 | Median gross rent as a percentage of household income | GDB |
| **2012** | acs5 | B25077 | Median Home Value: Owner-occupied housing units | GDB |
| **2011** | acs5 | B01003 | Total Population | NHGIS |
| **2011** | acs5 | B19013 | Median Household Income in 2011 | GDB |
| **2011** | acs5 | B19301 | Per Capita Income in 2011 | GDB |
| **2011** | acs5 | B23025 | Total Civilian labor force: Population 16 years and over | GDB |
| **2011** | acs5 | B23025 | Unemployed Population 16 Years and Over | GDB |
| **2011** | acs5 | B25001 | Total Housing Units | GDB |
| **2011** | acs5 | B25002 | Total Vacant Housing Units | GDB |
| **2011** | acs5 | B25058 | Median Contract Rent | NHGIS |

| | | | | |
|---|---|---|---|---|
| **2011** | acs5 | B25064 | Median gross rent | GDB |
| **2011** | acs5 | B25071 | Median gross rent as a percentage of household income | NHGIS |
| **2011** | acs5 | B25077 | Median Home Value: Owner-occu-pied housing units | GDB |
| **2010** | acs5 | B01003 | Total population | NHGIS |
| **2010** | acs5 | B19013 | Median Household Income in the Past 12 Months (in 2010 Inflation-Adjusted Dollars) | NHGIS |
| **2010** | acs5 | B19301 | Per Capita Income in the Past 12 Months (in 2010 Inflation-Adjusted Dollars) | NHGIS |
| **2010** | acs5 | | Total Civilian labor force: Popula-tion 16 years and over | imputed |
| **2010** | acs5 | | Unemployed Population 16 Years and Over | imputed |
| **2010** | acs5 | B25001 | Housing Units | NHGIS |
| **2010** | acs5 | B25004 | Vacant housing units: total | NHGIS |
| **2010** | acs5 | B25058 | Median Contract Rent (Dollars) | NHGIS |
| **2010** | acs5 | B25064 | Median Gross Rent (Dollars) | NHGIS |
| **2010** | acs5 | B25071 | Median Gross Rent as a Percent-age of Household Income in the Past 12 Months (Dollars) | NHGIS |
| **2010** | acs5 | B25077 | Median Value (Dollars) | NHGIS |
| **2000** | 2000_SF1b | FXS | Total Population | NHGIS |
| **2000** | 2000_SF3b | HF6 | Median Household Income in 1999 | NHGIS |
| **2000** | 2000_SF3b | HG4 | Per Capita Income in 1999 | NHGIS |
| **2000** | 2000_SF3b | HL0 | Total Civilian labor force: Popula-tion 16 years and over | NHGIS - constructed |
| **2000** | 2000_SF3b | HL0 | Unemployed Population 16 Years and Over | NHGIS - constructed |
| **2000** | 2000_SF1b | FV5 | Total Housing Units | NHGIS |
| **2000** | 2000_SF1b | FWB | Total Vacant Housing Units | NHGIS |
| **2000** | 2000_SF3b | G74 | Median Contract Rent | NHGIS |
| **2000** | 2000_SF3b | G8C | Median Gross Rent | NHGIS |
| **2000** | 2000_SF3b | G8L | Median Gross Rent as a Percent-age of Household Income in 1999 | NHGIS |
| **2000** | 2000_SF3b | G8V | Median Value | NHGIS |
| **1990** | 1990_STF1 | ET1 | Total Population | NHGIS |
| **1990** | 1990_STF3 | E4U | Median Household Income in 1989 | NHGIS |
| **1990** | 1990_STF3 | E01 | Per Capita Income in 1989 | NHGIS |
| **1990** | 1990_STF3 | E4I | Total Civilian labor force: Popula-tion 16 years and over | NHGIS - constructed |
| **1990** | 1990_STF3 | E4I | Unemployed Population 16 Years and Over | NHGIS - constructed |
| **1990** | 1990_STF1 | ESA | Total Housing Units | NHGIS |
| **1990** | 1990_STF1 | ETQ | Vacant housing units | NHGIS |

| | | | | |
|---|---|---|---|---|
| **1990** | 1990_STF3 | ES6 | Median Contract Rent: Specified renter-occupied housing units paying cash rent | NHGIS |
| **1990** | 1990_STF3 | EYU | Median Gross Rent: Specified renter-occupied housing units paying cash rent | NHGIS |
| **1990** | 1990_STF3 | EY3 | Median Gross Rent as a Percentage of Household Income in 1989: Specified renter-occupied housing units paying cash rent | NHGIS |
| **1990** | 1990_STF1 | EST | Median Value: Specified owner-occupied housing units | NHGIS |