A Tool to Visually Analyze Homelessness Trends in U.S. Metros

Lawrie Brunswick, Manya Chadha, Parvati Jayakumar, Sarah Nguyen, Jacob Peterson, Kyle Sorstokke, Daniel Vogler

View our visualization here

Part I: Executive Summary

This report outlines a tool developed to empower researchers, non-governmental organizations (NGOs), and local policymakers with comprehensive visual insights into homelessness trends in US Metropolitan Statistical Areas[1] (MSAs) since 2007. The tool aims to enhance decision-making and resource allocation in addressing challenges related to homelessness through various associated factors.

Key Findings

1. Current Challenges

- Homelessness is idiosyncratic with a mixture of factors likely contributing to the outcome; depending on the city, different factors are associated with it.
- Current homelessness data is decentralized, as governmental and non-governmental agencies collect and report in numerous unconnected ways.

2. Data Visualization and Integration

- Aggregates data from 6 government agencies[2] responsible for data such as economic, health, housing, and demographic trends.
- Interactive visualizations allow users to quickly build national and city homelessness outlooks and direct resources accordingly.

3. Our Homelessness Trends Tool

- Introduces a user-friendly web-based dashboard providing extensive data from 2007-2020 for roughly 50 largest U.S. metro areas on geographic distribution, demographics, and emerging patterns in homelessness rates.
- Utilizes advanced analytics to identify city-specific key factors contributing to homelessness.

Conclusion

The homelessness trends tool is a convenient aide to researchers and policymakers working to address homelessness by understanding recent trends. The tool equips stakeholders with the data needed for targeted interventions in their cities. Collaboration between organizations will have a positive impact on the homeless population in large U.S. metro areas.

---

[1] Most policymakers are responsible for addressing homelessness locally, not in the country as a whole.
[2] Department of Housing and Urban Development (HUD), Federal Reserve Economic Data (FRED) service from the Federal Reserve Bank of St. Louis, Federal Housing Finance Agency, Centers for Disease Control (CDC), U.S. Census Bureau, Bureau of Economic Analysis (Department of Commerce)

Part II: Concept Background

Background Research

Academic research surveys[3] reveal that communities with higher unemployment rates, limited job opportunities, or lower income levels experience increased homeless rates. Areas with limited affordable housing options, lack of strong community networks and social welfare programs follow similar trends. Moreover, demographics distribution of race, age, and gender within communities can impact vulnerability of certain groups to homelessness. The causes and characteristics of homelessness are numerous, and to ensure an appropriate scope, we focused on replicating established themes in the literature[4] that identified variables that are predictive of homelessness, within metro cities specifically:

- **Rental Market Variables** such as low-cost rental housing and house price index, which measures the change in value of single family homes
- **Macroeconomic Indicators** such as inflation and unemployment
- **Public health factors** such as the number of those uninsured or with a disability, and death by unintentional drug overdose and alcohol consumption
- **Strength of safety net** programs such as the number of beds available within homeless shelters and availability of welfare benefits such as the Supplemental Nutrition Assistance Program (SNAP), which are offered to those who fall under a certain threshold of the poverty line

In addition to these potentially causal factors, we collected specific information about how the homeless population breaks down according to various **demographic characteristics** such as age, race, gender distribution. Our tool makes sense breakdowns easily visually accessible for each US MSA with sufficient data.

---

[3] See, for example:
Fargo JD, Munley EA, Byrne TH, Montgomery AE, Culhane DP. Community-level characteristics associated with variation in rates of homelessness among families and single adults. Am J Public Health. 2013 Dec;103 Suppl 2(Suppl 2):S340-7. DOI: 10.2105/AJPH.2013.301619.

[4] Deden Rukmana (2020) The Causes of Homelessness and the Characteristics Associated With High Risk of Homelessness: A Review of Intercity and Intracity Homelessness Data, Housing Policy Debate, 30:2, 291-308, DOI: 10.1080/10511482.2019.1684334,

Data Handling - Acquisition & Profiling

To make our integrated dataset which encompassed these variables, we gathered and compiled data from six different sources *(Figure 1).* As most of our data is quantitative, we had to carefully prioritize how to encode each variable, ensuring the most important ones are most prominent. Homeless population was given the highest priority as we wanted to highlight how the rate of homelessness (calculated per 100,000 people) changed once plotted against each of the other variables.

Source	Variable	Data Type
U.S. The Department of Housing and Urban Development (HUD)	1. Total Homeless Population 2. Demographics including race, ethnicity, gender, and age 3. 1-Bedroom Rentals	1. Ratio 2. Ratio 3. Ratio
U.S. Census Bureau	1. Number of Beds 2. SNAP recipients 3. Population of uninsured people 4. Population with a Disability	1. Ratio 2. Ratio 3. Ratio 4. Ratio
Bureau of Economic Analysis (BEA), U.S. Department of Commerce	1. Total Population	1. Ratio
Federal Reserve Bank of St. Louis	1. Inflation 2. Unemployment	1. Ratio 2. Ratio
Federal Housing Finance Agency	1. House Price Index	1. Ratio
Center for Disease Control (CDC) and Prevention	1. Unintentional Overdose Deaths 2. Death due to Alcohol Consumption	1. Ratio 2. Ratio
—	1. Metropolitan Statistical Areas	1. Nominal
—	1. Year	1. Interval

Figure 1: Where Our Data Comes From

Data Preparation

Bringing together datasets from various places required preparation and cleaning. One of the primary challenges of this was the difficulty of merging Metropolitan Statistical Areas data included:

- Each datasets had different variations of how they labeled their MSA
- Some did not have data at that granularity, or used county level data instead.

Due to this, we filtered our data to only keep information about the largest 50 MSAs and manually mapped county level data to each MSA.

Once each member finished compiling their respective datasets by keeping only relevant columns in a standardized format, we merged all of our clean data together using Google Colab. To ensure the merging data from all six sources was seamless, unit tests were implemented to check that each dataset had an ‘MSA’ and ‘Year’ column, and that there was no duplicate data for these columns. Once all datasets passed the sanity check, the datasets were merged to create one large dataset.

Part III: Process Description

A Timeline of Guerrilla Usability Test Results

The usability timeline encapsulates the dynamic journey of adapting the dashboard to meet user needs, ensuring a more effective and user-centric final product. We conducted usability testing to assess the user experience of the visualization concept and design, focusing on key aspects such as feature navigation, functionality, comprehensibility and overall user satisfaction. A diverse group of participants provided insights into usability across different skill levels *(Figure 2)*.

The first phase involved peer user testing using a paper/low-fidelity prototype. Next, we conducted expert usability tests from industry professionals and a professional graduate student with wireframe navigation pages using pseudo data. Lastly, we demonstrated the homeless dashboard tool in its final stages to more peer users.

Participant User Profiles

Name	User Persona/Job Title	User Story
Ahana Raina	Grad Student - MPA (Public administration)	Assess impact of different policies (public health, macroeconomic, etc.) on rates of homelessness.
Lisa Renfro	County Level: Manager of Social Services	Get a quick overview of the state of homelessness and resources available. Build case studies of success and failures when planning next steps at the county level.
Nancy Aguirre	County Level: Research Specialist II	To develop useful reports for stakeholders and managers involved in development of local pilot programs aimed at improving homelessness
Vanja Glisic	MSDS grad student / Data Analyst at Liberty Mutual insurance	Predict future trends in homelessness by forecasting related variables like median salaries
Chakita Muttaraju	MSDS grad student / ex-SDE at Amazon	Get introduced and gain high-level insights to the topic of homelessness in the U.S.
Apoorva Sheera	MSDS grad student / ex-consultant at McKinsey	Assess trends in homelessness and relate them to trends in other variables like housing prices and inflation.

Figure 2: Expert and Peer User Profiles

The iterative design process is evident through the incorporation of user suggestions and the commitment to addressing concerns for a more user-centric and effective dashboard. We measured user feedback in an open-ended manner, guided by a user satisfaction survey. Feedback from peer and expert users highlights the importance of clarity, precision, and user-friendly features in the dashboard design. Although feedback regarding the visual design of the interface was generally positive, there were suggestions for improvement in the fine details such as:

- Participants expressed difficulty in locating specific features and information.
    - The US map dominates user focus on the homepage which causes users to overlook line charts in the side panel.
- Several users suggested a more intuitive predictive variables page layout, emphasizing the importance of semantic understanding of measures to gain usable insights.
    - Nancy Aguirre commented, “Just remember that your audience does not understand the data the same way you do”.

After phase I and II of usability testing, we addressed concerns by adapting navigation configurations. We paid close attention to overcrowding to minimize the occurrence of user oversight. The homepage now allows users to select an MSA on the US map and navigate to a summary of all predictive variables for that MSA and visualize demographics for the homeless population. We also added hue encoding for correlation values to give priority to predictive variables in an intuitive way.

In the final phase of testing, we performed a live demonstration of the final draft of the dashboard. During this demo we walked through such features as the drilling down from the homeless map to predictive variable comparisons for a particular MSA. After the demo, users requested features such as:

1. A chronological view of predictive variables
2. Sharper color contrast on the map
3. More labels for trend lines and confidence intervals
4. An indication that homeless is size and hue encoded on the map
5. An animation on the map that loops through different years
6. Captions explaining features of predictor variables

After phase III of usability testing, we tailored our final edits with these comments in mind.. We updated the detailed scatter plot pages to include a time series view of predictive variables and homeless rate on a double y-axis. We used red bubbles on the map to give a sharper contrast and improved intuition by carrying the red hue to all pages indicating the homeless rate. We moved the year dimension to pages instead of filters. This allowed us to animate the visualization and loop through the homeless rate for each year. Lastly we added captions to the bottom of detailed predictor pages, for a better understanding of measure definitions.

We determined that a few feedback suggestions were out of scope. As a manager of social services, Lisa Renfro wanted to see the availability of housing in the median rent range and options for low-income households. In addition, several users wanted to see and compare geo-encoding predictor variables along with the homeless rate. Unfortunately, these outlying requests were not feasible in the time we had.

Problems Encountered and How We Addressed Them

As we produced this dashboard, we encountered challenges that can be grouped into three areas: limitations of the data itself; challenges constructing a unified dataset from the disparate sources of data; and challenges producing the final visualization.

Limitations of Datasets

The translation of qualitative descriptors of homelessness into specific, observable quantitative variables that serve as a meaningful proxy for that factor.

Example: Although it may seem obvious that “high housing prices” in an area, keeping everything else constant, would make housing every member of that area more challenging. However, measuring this effect quantitatively is another question. There are many data sources we can use to quantify “high housing prices”. We can report both rental and purchase prices for various sizes of homes. Ultimately, this requires us to use our judgment and contextual understanding to filter the possible data sources and identify the proxy variables that are most likely to be reliable. For this reason, we deferred to government sources whenever feasible (i.e., whenever doing so allowed us to get datasets that were as complete as possible and included the targeted variables).

Limitations Merging Datasets

A second category of challenges we experienced did not have to do with any one of the datasets per se. Instead, it has to do with combining datasets collected in different contexts in a way that is likely to be useful to people studying homelessness.

Example: One issue was standardizing place names and dataset formats. Our workflow assigned each member of the team a specific dataset to find, clean, and preprocess. To complete this task seamlessly as possible, we created a standardized data schema specifying the columns each dataset should contain once preprocessing was complete. Even after taking this step, we encountered standard issues with slightly different ways of representing the same place (“Oakland, CA” vs. “Oakland, California”, etc.). We learned that a successful data schema specifies even the most mundane details and conventions precisely to avoid inconsistencies that make merging different datasets challenging.

Example: Another major challenge was that some of our different datasets were organized by geographic units that were not consistent with each other. More concretely, some data were reported by each county; other data were reported by each one of the HUD’s continuums of care; other data were reported by MSA. This required us to map and sum them to produce a dataset organized by consistent geographical units (MSAs). Sometimes, standards and definitions of these geographic units changed over time, requiring us to make adjustments to keep the data consistently organized (“Seattle-Tacoma-Bellevue MSA” vs. “Seattle-Bellevue-Everett MSA”).

Example: Finally, not all data we were interested in was available for every city and every year within our scope of study. This required another choice: build visualizations based on the most comprehensive dataset possible, which would have no data available for some city-years, or visualize the ‘lowest common denominator’ of city-years with full data availability. Ultimately, we decided the best choice was to publish everything, including city-years with no data, and develop a user interface that transparently alerted users to lacking data.

Visualization Challenges

Beyond the underlying datasets, we had another set of difficulties to overcome related to creating the visual itself. These can be further segmented into:

1. Design issues: Iterating on our original design ideas after realizing that the draft visualizations we produced were not as effective as anticipated while brainstorming
2. Process issues: Translating our mock-ups into effective visualizations in Tableau.

We sometimes succeeded technically in building out features that did not turn out to be as visually effective as we had anticipated during the design phase, requiring iteration. For example, we produced visuals allowing users to view lagged correlations between predictor variables and the homelessness rate. In other words, if we plot the homelessness rate in a city in year *t*, users could see how closely correlated that was to indicator variables in years *t-1, t-2,* up until *t-5*. However, during our presentation we realized that (a) that users found it confusing and (b) *generally*, the strongest correlations were between homelessness in year *t* and predictor variables in year *t*, which led us to drop the visualization altogether.

Even after adjusting our designs in response to the effectiveness of our prototypes, we encountered challenges in the process of translating those designs into effective final visualizations. Some related to our choice of Tableau as the primary visualization tool, which required making trade-offs compared to alternatives: Tableau's drag-and-drop interface makes exploring our data easy, but also means documentation of steps leading to a visualization is not built into the workflow in the same way it is in code-based alternatives. This sometimes creates problems reproducing visuals. This was relevant because another bottleneck was the inability for all of our team members to work in one Tableau environment at the same time, requiring us to work separately then try to reproduce the visualizations.

Insights and Breakthroughs

Relationships

Our insights and breakthroughs center primarily around the research questions with which we began. Some of the most prominent include the surprising differences between certain cities, trends that hold only in specific geographic regions, and clear relationships between certain predictors and homelessness. While it is reasonable to assume that most large metro cities would show similar trends, we found an unexpectedly large difference in some. Nearly all of the factors in question had their homelessness correlation fluctuate throughout the list of cities, with many split between positive and negative. ***One of the most consistent relationships was the strong positive correlation between the number of shelter beds available and the homelessness rate***. While we primarily set out to find factors that impact homelessness, this insight is likely the opposite; cities facing higher rates of homelessness often expand the availability of safety net programs.

Regional Trends

Many of the insights discovered only apply in specific regions of the country, such as the percentage of the population without insurance in Southern and Midwestern cities or the exponential increase in housing prices in the Bay Area. The cities with positive correlations between uninsured percentage and homelessness rate are clustered largely in the South and the Midwest, with opposite trends appearing in cities like Los Angeles and New York. Further research is required for confirmation, but this could be influenced by different state governments' handling of the Affordable Care Act and the political push for wider healthcare coverage. As far as housing prices, most cities showed a somewhat negative correlation to homelessness; however, San Francisco, San Jose, and Sacramento behaved oppositely. The Bay Area is known for struggling to lower homelessness rates, as well as for the rapid increases in housing prices and overall cost of living related to the growth of its tech industry. Since these two factors both stand out from national trends in the region, it is understandable that the relationship between them may also be extreme.

Technical Process

We also reached valuable breakthroughs in the technical process throughout this project, such as clearer and more legible visualization techniques and preparing large disparate datasets. One visualization insight was the constraints imposed on presentation of chronological data; our initial attempts failed to convey information as we intended because the chronology was not fully handled. Incorporating a time series alongside the plot of observations helped filter out the noise of cyclical variable changes to focus on the true underlying relationships. The biggest technical breakthrough surrounded data cleaning and merging. With so many different data sources, all providing large and often uncleaned data, properly formatting and joining our final dataset was a huge task. The insight that provided a path forward was the development of a clear structure for each subset and performing chunks of the work in the language or IDE that was most effective for each person. By putting clear constraints on the desired format of each data source, we were able to more efficiently clean them with a combination of Python, Tableau, and R before the simple final merge.

Part IV: Critical Evaluation

Our project aimed to effectively convey insights from the integrated homelessness dataset, employing various visualization techniques and concepts learned in the course. Nominal comparisons, time series, ranking, part-to-whole, deviation, distribution, and correlation were effectively represented using appropriate visualizations such as stacked bar charts, horizontal bar charts, correlation maps, choropleth maps, and time series line charts.

The **Principle of Consistency**[5] was adhered to, ensuring that the properties of visualizations matched the properties of the data, i.e., quantitative data was encoded using consistent shape and length, unnecessary 2D and 3D encodings were avoided for 1D data to maintain simplicity,

[5] Mannheimer, Nathan, "Graphical Excellence and Integrity", Data 511, Slide 19

values were adjusted consistently for population changes when applicable, and labels were applied carefully and consistently for clarity and understanding.

The visualizations demonstrate expressiveness by presenting the set of facts accurately (as per **Mackinlay's Design Criteria**)[6]. The effectiveness of each visualization was evident through the ease with which information could be perceived. The principle of maximizing the **"Data-Ink Ratio"**[7] was consistently applied, emphasizing relevant data while eliminating non-essential ink. This principle was especially evident in the careful labeling, and avoidance of unnecessary 2D and 3D encodings.

Our project also tackled the challenge of selecting the "best" visual encoding for a given set of data variables. The **Principle of Importance Ordering**[8] guided the process, ensuring that the most critical information was encoded effectively. For instance, time series data was presented using line charts, emphasizing the value at each point in time.

Our in-class discussions on the effectiveness of **bar charts versus pie charts**[9] and the use of stacked bar charts guided our thought process behind choosing specific visualizations. Our project acknowledged the difficulties in assigning quantitative values to pie chart slices and the challenges of comparing stacked bars due to the need for mathematical interpretation.

The use of choropleth maps effectively conveyed geographical variations in the homelessness data. Our project considered size over color as the encoding attribute, recognizing its effectiveness for **representing quantitative values**[10] on maps.

We have also emphasized the need to understand the domain and collection methodology, check data ranges, and document the cleaning process. The focus on handling outliers based on the analysis's purpose demonstrated a nuanced understanding of data quality.

Our project incorporated multivariate analysis through interactive visualizations and dashboards. We have recognized the importance of **interlocking feedback loops**[11], allowing users to manipulate, explore, and solve problems interactively. The inclusion of diverse visualizations in the dashboard facilitated strategic, analytical, and operational tasks.

Overall, we have met the project’s goals by incorporating a range of visualization techniques and concepts. The critical evaluation underscores the thoughtful application of principles such as consistency, importance ordering, and maximizing data-ink ratio, contributing to the project's effectiveness in conveying complex insights from the dataset.

---

[6] Mannheimer, Nathan, "Graphical Excellence and Integrity", Data 511, Slide 28
[7] Mannheimer, Nathan, "Graphical Excellence and Integrity", Data 511, Slide 46
[8] Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten.
[9] Mannheimer, Nathan, "Graphical Excellence and Integrity", Data 511, Slide 103
[10] Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis.
[11] Mannheimer, Nathan, "Multivariate Analysis, Interaction Techniques, UCD", Data 511, Slides 142, 143

Appendix

Exhibit 1: User Experience Questions

- How approachable is the prediction oriented setup with staggered years?
- Is any part of the dashboard navigation confusing or unintuitive?
- For what purpose are you most likely to use this dashboard?
- Is there anything left unanswered that you wish the dashboard included?
- Which feature or visualization felt the most useful for your specific user needs?
- How could this impact your effectiveness on the job?
- What was easy or difficult about navigating the dashboard?
- How simple and clean was the interface layout?
- Which parts of the dashboard would you use most often?
- What did you think of the explanations and titles on the page?
- What do you think [feature] is trying to communicate to you?
- Are there any important questions that are left unanswered after using the dashboard?

Exhibit 2: Prototype Figures

Disclaimer: prototype features were **illustrative, not actual data,** and used for test purposes

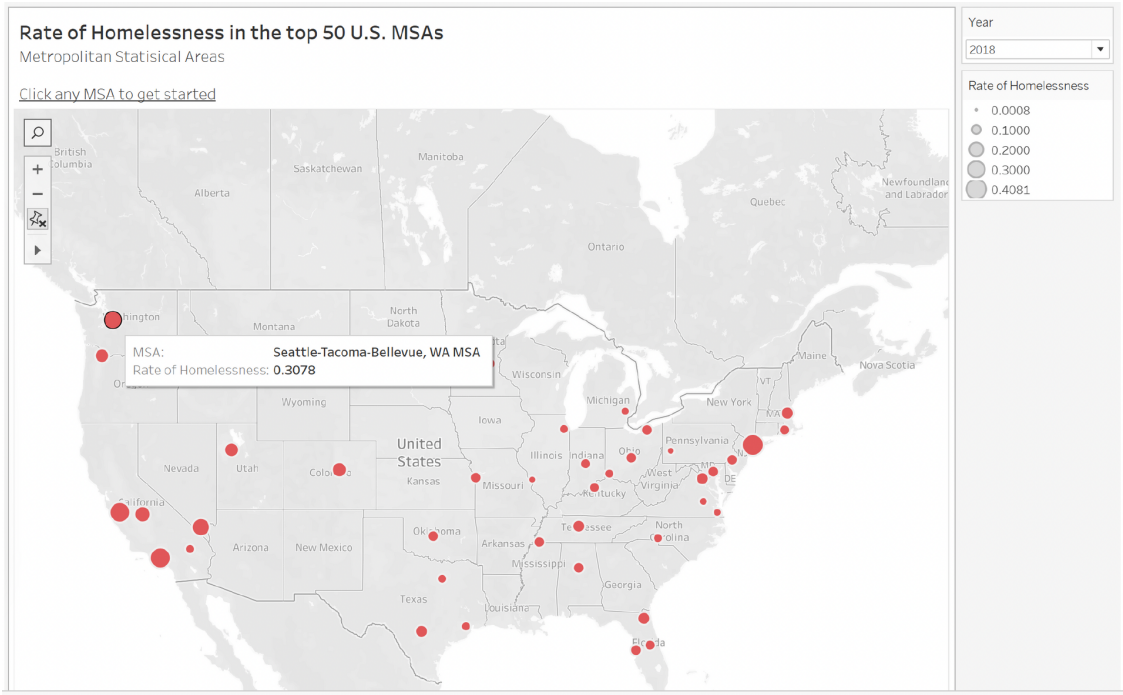

Figure 3: Mock Dashboard - Map

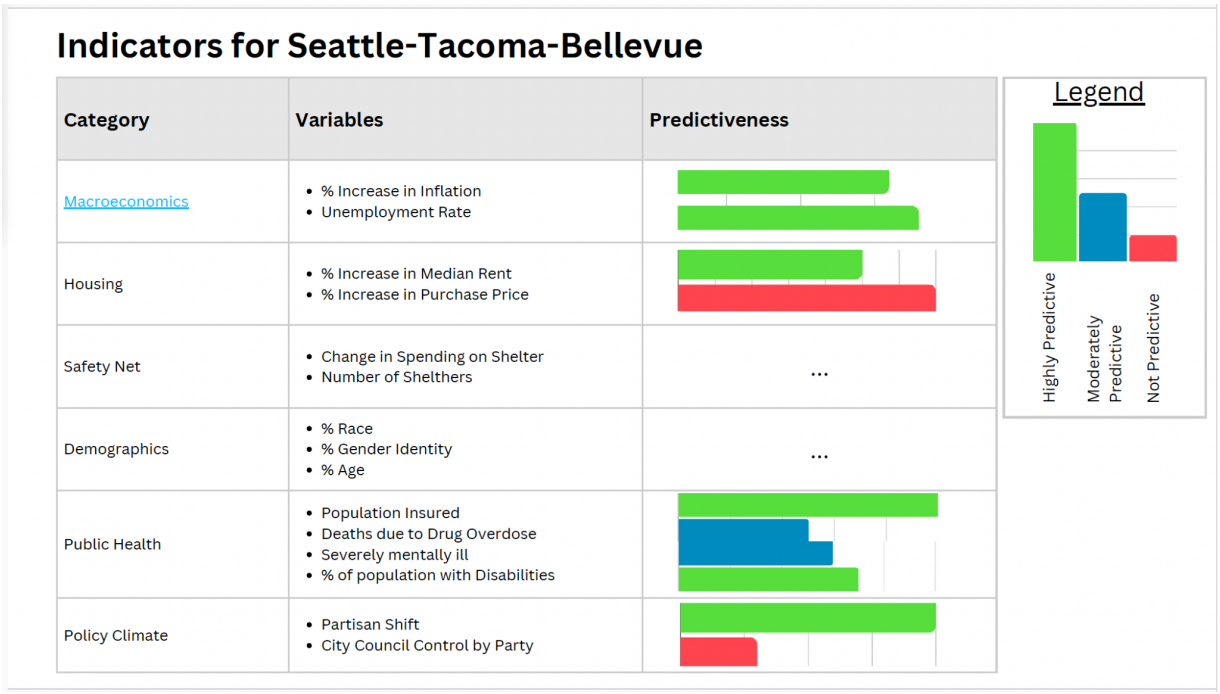

Figure 4: Mock Dashboard - Predictor Navigation

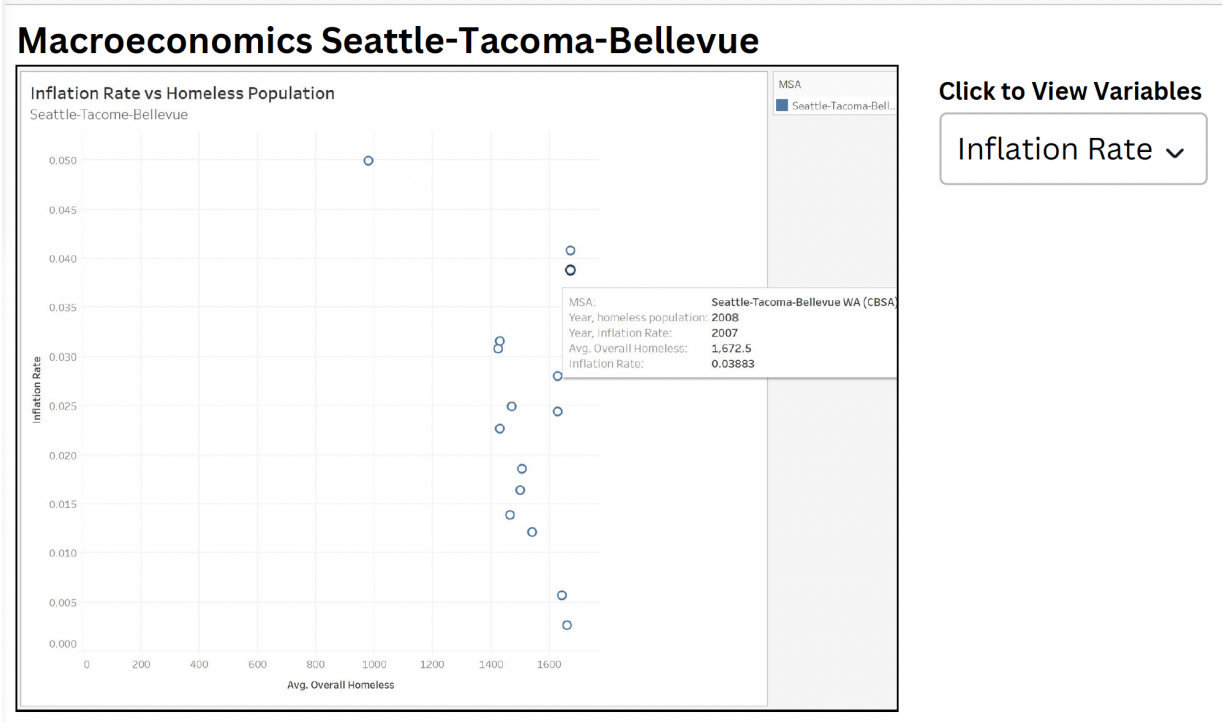

Figure 5: Mock Dashboard - Details On Demand In Feature

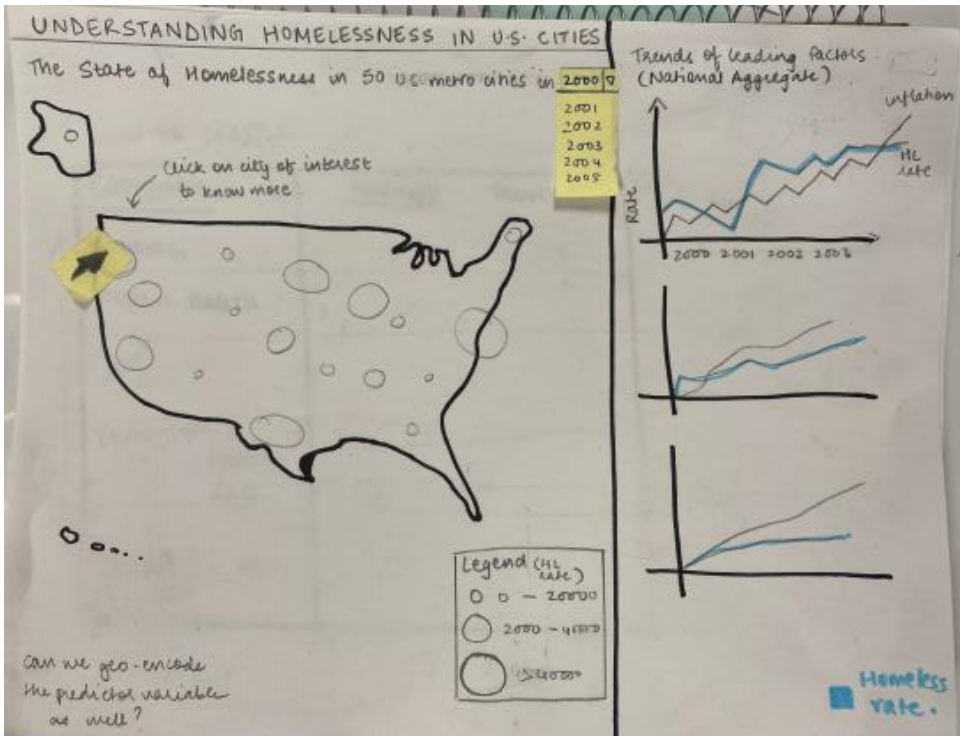

Figure 6: Low-fidelity prototype - Map

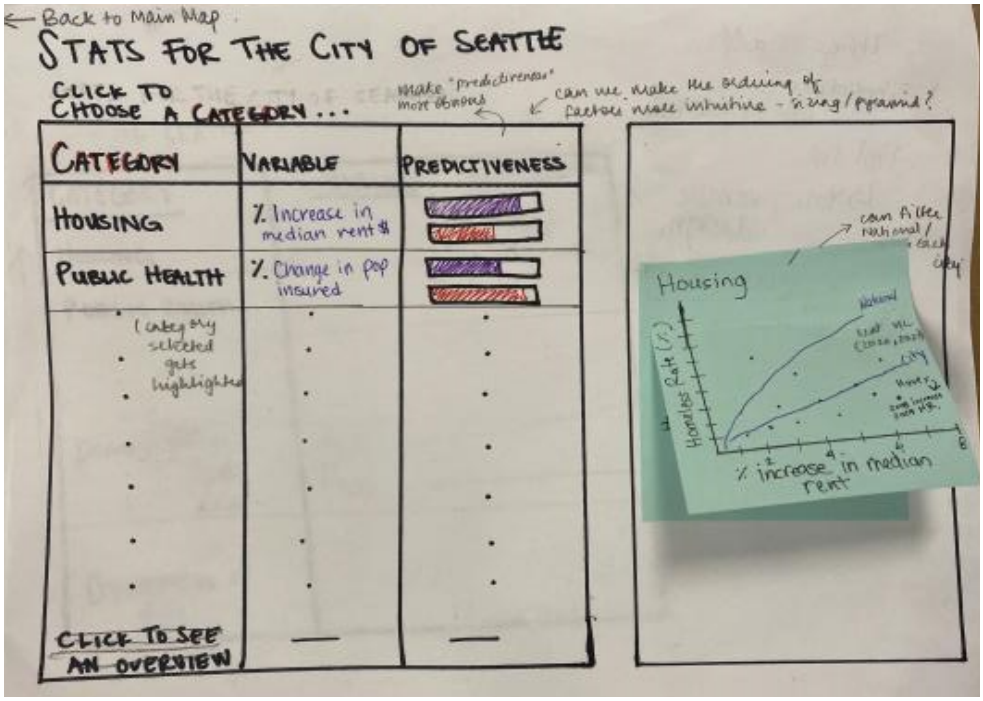

Figure 7: Low-fidelity prototype - Predictor Table