Forecasting Aggregate National Data with Interactive Visual

Team 30
Amandeep Singh, Kyle Lindteigen, James Renier Domingo, Yogesh Raparia, Michael Jones, Paul Horton

What

Interactive tool that shows users key metrics of US communities including:

- Population
- Income
- Housing Price Index
- Temperature
- Road Safety
- Job Growth

Why

US demographics are rapidly changing. Anyone using these metrics will need to understand recent trends and projections. This tool could be useful for the following cases:

1. Local governments seeking to impove
2. Individuals looking to move
3. People wanting to better understand their community

How is rent expected to change to change in Georgia vs. Texas over the next 5 years?

Does California or Washington have a lower crime rate?

How

Data

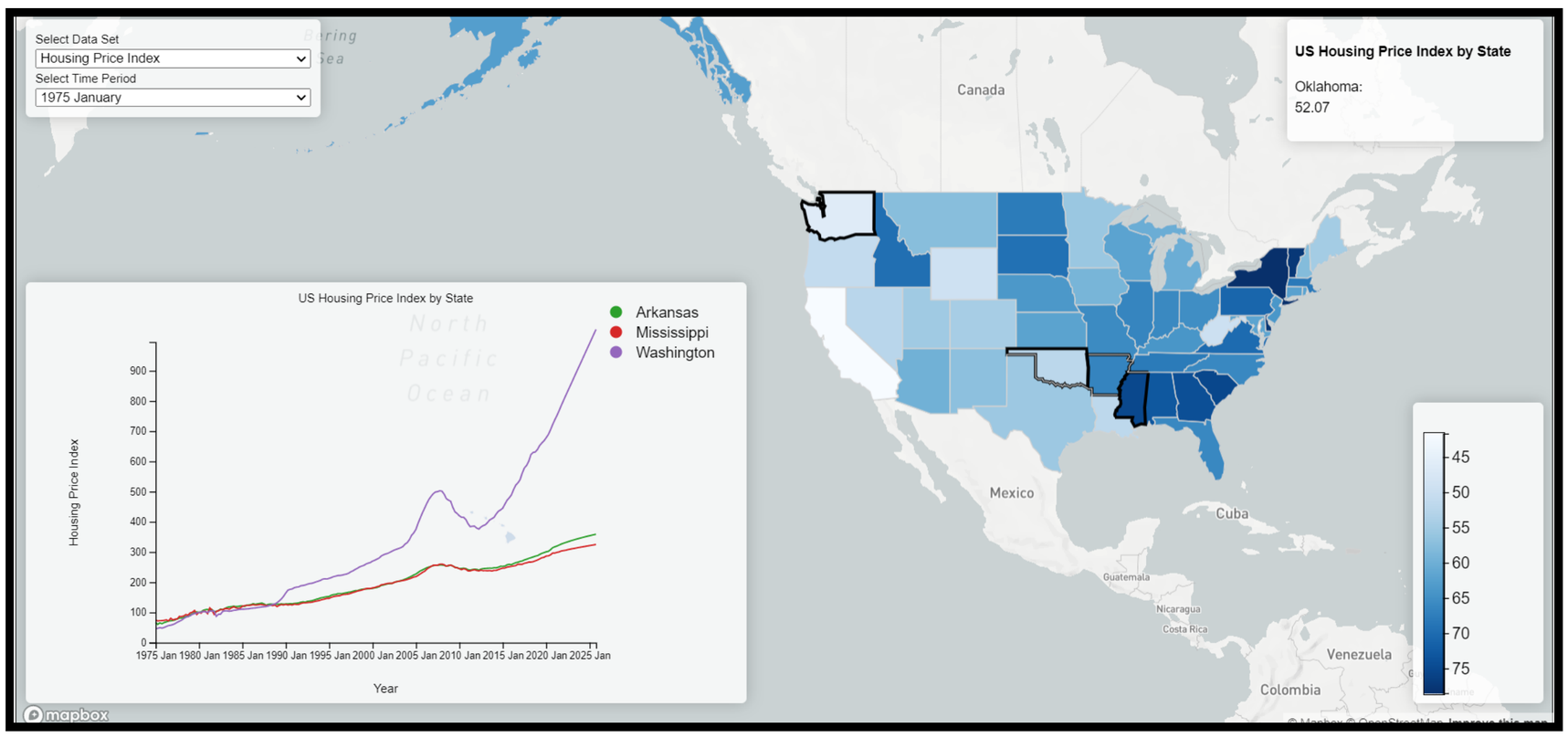
Data collection and cleaning were significant challenges for this project due to the decentralized nature of demographic data. Our data come from the Census Bureau, Bureau of Economic Analyses, National Oceanic and Atmospheric Adminsistration, the FRED, and the Bureau of Justice Statistics. Additionally, these organizations have varying standards and formats. Some aggregate by state and others county. They do not have the same time periods with some observations every month and others every quarter. The temperature data included observations dating back to 1895 for each county so this dataset had almost 400,000 observations alone. Other datasets were structured by county information, quarterly observations, or seperated by ethnicity and age. Each dataset was individualy standardized to an NxD format where N represents the observation time and D represents the state.

Models

Our research indicated two types of models that proved successful for predicting and forecasting new data:

- ARIMA models fit to time series data and forecast futures values based on past data. Parameters were optimized using cross validation based on the minimum AIC to determine the model's.
- Random forest models aggregrate decision trees to predict values and variable importance. We tuned the number of trees, node size, and variable usage parameters with cross validation.

Tool



Features

1. Dropdown menu for variable selection and time period
2. Choropleth map of the US
3. Continuous scale indicating variable value
4. Tooltip which displays value for state with mouse hover
5. Line chart with trends and projections for selected variable
6. Only selected states appear on the line graph and show with an outline on the map

Results

The evaluation of our tool consists of two parts: accuracy of predictions and usefulness of the tool. We have established separate methods for the evaluation of our performance for these two aspects.

For the prediction accuracy, we have divided our data into train and test sets. The ARIMA models use the last 5 years as the test set with all other years as the training set. Our table with results for Root Mean Squared Error on all predicted variables is below. The Random Forest model indicated through the variable importance output that the presence of a traffic signal was the most significant variable for accident severity in the majority of states. Our results are considered reliable after comparison from similar models in our literature survey. State trends for 2020 differed markedly from previous years so the pandemic has created significant amounts of error for our models.

We created a Google survey to gather feedback about the user experience when engaging with our tool. Despite having many potential uses, we told the users to consider themselves as looking for a new location to live. We found that there are opportunities for improvement but the feedback was generally positive. Chart 1 to the right shows response distributions for one section of questions on the tool's features.

Metric	Population	Income	Temperature	Housing Price Index	Jobs Added
Mean	6,366,000	59,600	64.1	418.6	183.9
RMSE	1,323,000	6,230	1.45	5.7	0.7

Innovations

Users can compare many relevant community factors with a single tool. The tool combines historical values with predicted trends allowing users to see how values will change in the future. Our visualization includes time-series and geographic data with interactive features to give users more ability to explore and deliver insights.

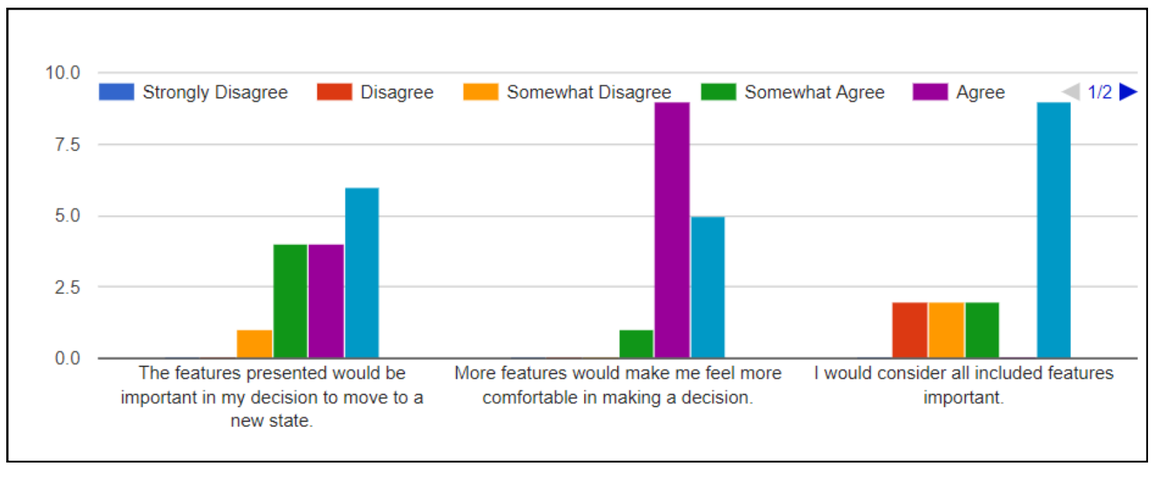


Chart 1: Resposes for three questions regarding the features present in out visualization tool.