Market Maven

Daniel Bae, Keith Dowd, Max Smoller, and Ruoyi Chen

1 Introduction and Problem Definition

Keeping up with the complex network of factors that influence the U.S. real estate market is a challenge, even for professional investors. While online tools, such as *Zillow*, facilitate market research and comparison, they are limited in scope and functionality, as these tools are primarily focused on searching individual property listings. We believe there is a large gap between the level of market research capable with current tools and the demand for higher-level market research capabilities at the macro level. This gap between supply and demand is the primary motivating factor for developing a new tool to meet this demand. We set out to create a data-driven tool with the goal of addressing these gaps in the market research landscape.

Market Maven is a novel web application for users to explore and compare real estate markets. By leveraging advanced analytics and visualizations, users are empowered to learn about different markets across the U.S. via a wide breadth of characteristics. This functionality is made available within a single, user-friendly web application. The audience for *Market Maven* includes home buyers, sellers, and investors who want to better understand and compare real estate markets. Individuals who are looking to move to a new area of the U.S. may find this tool especially valuable. State and city planners may also find it useful for decision-making related to local development and resource allocation.

Market Maven is available now and freely accessible online at https://www.mymarketmaven.com.

2 Survey

Market Maven builds on existing bodies of work by expanding scope and providing information in an easier to digest format. The following are some areas of overlap with existing works.

2.1 Real Estate Market Factors

There is a great amount of literature around exploring different factors that affect the real estate market. Some of these factors include creative subcultures attracting high-tech jobs to specific areas of the country [1], differences in regulations and available land affecting cities' housing supply [2], and how companies are reshaping their work space to utilize less office space [3]. These insights should be useful as they shed light into some influencing factors that cause differences in real estate markets and key demographic and economic indicators in regions across the US. It is important for our tool to highlight these and similar factors which can be used to draw comparisons between markets. Many of these factors can also be used to predict future market outlook and can be an influencing factor in a clustering algorithm. The above features do not encompass all possible influencing factors that describe and differentiate real estate markets and further research and market analysis will need to be done to explore these missing components.

2.2 COVID-19

The COVID-19 pandemic has caused a sharp contraction in the real estate market and housing prices [4] [5] [6]. There may be great interest to compare housing and rental prices before and after the pandemic. It will also be of interest to determine whether real estate prices can recover after this pandemic and how this affects key real estate and demographic metrics. For this reason, it is crucial to have full disclosure within *Market Maven* and display the time range for the data, as there is likely a clear distinction between pre- and post-COVID-19 data. All data shown in *Market Maven* is data collected before COVID-19.

2.3 Existing Applications

Several applications facilitate the exploration of real estate markets. Geography, human mobility, and transaction data were used to create a system for discovering high-value markets [7]. While this tool used statistical models and diverse data, it was not accessible online and focused only on a single market dimension. This tool is less useful for *Market Maven* because it looked only at high-value markets. In another example, a web-based tool was developed that used modern visualization techniques, including stacked graphs and clustered pixel bars, to show sales over time [8]. User research found these visualizations were helpful for home buyers and investors, alike, to learn about markets; however, other market characteristics were not considered. These visualizations are useful to consider for inclusion in *Market Maven. HomeSeeker* improved on these tools by reporting a wider array of data, such as property and retail, and presented these data using modern visualizers like choropleth maps [9]. Logical categories (e.g. economy, social, etc.) for market data were also provided, which will be useful for selecting data to include in *Market Maven*. These efforts are ripe for improvement. No effort allows for comparison of multiple markets. While a diverse set of data is used across these tools, none blends all these data into a single tool.

2.4 Geography and Environment

A study suggests that the economic geography of the U.S. may be better understood within the context of *mega-regions* [10]. Furthermore, areas characterized as urban, suburban, and rural all have unique traits and have undergone very different changes over time [11]. The main opportunities for improvement on the existing research is to expand the scope beyond commuter data and implement more robust analytic techniques.

Research suggests that due to climate change, particularly in vulnerable areas such as coastal regions, a growing dichotomy between supply and government subsidized demand has led to policy driven inflation in prices [12]. Though this research focuses on storm severity and encroaching shorelines in coastal regions, it may help us consider the importance of environmental factors in our own research on property value. This is also a key opportunity for improvement and generalization as the current research is highly focused on the north east coastal region of the U.S.

3 Proposed Method

3.1 Innovation

Several innovations differentiate *Market Maven* from other market research tools. Users are able to explore a wide range of real estate market metrics, including economics, demographics, weather, and more. Our comparison feature allows users to directly compare different counties to find similarities and differences. The tool also features a clustering algorithm that returns the most similar counties relative to a selected county. This feature opens up a recursive option for learning about and discovering new regions of the country. Such functionality is not available today given the variety of metrics we are incorporating, as well as the novel methods we employ to combine them. Finally, our interface is presented in a clean and intuitive way that grabs the user's attention without inundating them with all the information at once. In the following section we describe the specific innovations that we implemented within *Market Maven*.

3.2 Functionality

A choropleth map of the U.S. serves as *Market Maven's* landing page and visualizes real estate markets by county. An interactive dropdown allows users to select a metric they want to visualize on the map, including metrics that no existing real estate tool provides, such as market-level data for education levels, housing occupancy percentages, and economic indicators. A single color palette was used to visualize most of the metrics on the map. This was in order to simplify the user experience and allow for the user to focus on the information presented. However, for two of the metrics that included both positive and negative values, a diverging color palette was used to adhere to visualization best practices. The scales used to define the transition from one color to the next across the color palette were curated by metric to allow for effective visual differentiation between different values. Curation of the color scales enables the user to more easily spot differences and find locations of interest. Users can interact with the map by panning, zooming, and selecting a county for a more detailed inspection. After selecting a county, the user is provided with a pane where they can view a quick snapshot of the county's full suite of metrics, with a navigation link to the details page for that county.

The details page presents data in a dashboard format and allows the user to focus their review on a selected county. A wide range of metrics are provided to the user and are grouped into logical categories, such as economy, social, and geography. Percentiles are displayed to the right of each metric to put it into perspective relative to other counties. To facilitate additional comparison, the user is also able to select a second county to compare with, and the metrics and percentiles for the comparison county are displayed side-by-side. A list of the top-5 counties most similar to the selected county is included at the bottom of the details page. A link is provided to allow users to easily navigate to the clustering page to continue their exploration. Ultimately, the details page provides users with a view of the data that encourages iterative data discovery. A user can select a county and explore its metrics. They can select a comparison county to easily spot similarities and differences. Then, they can easily identify similar counties to the one selected and explore its metrics, too. This cycle encourages and facilitates new discoveries.

The clustering page presents an interactive visualization for the user to find counties similar to their selected county. All U.S. counties are displayed on the visualization and their location is based on their similarity to one another, where counties closer in proximity are more similar. Colored diamonds represent counties. On the default screen, the colors represent each cluster. The user can interact with the visualization by zooming, panning, hovering, and clicking on the colored diamonds. Dropdowns are also provided for the user to select a county of interest. Upon selecting a county, the user is presented with a view that shows their selected county and its thirty most similar counties. Once a county is selected, colors are used to represent proximity. A difference score is shown when a user hovers over a county. This difference score is a numeric representation of the similarity between the selected county and the hovered over county. The larger the difference score, the less similar the hovered over county is to the selected county.

Finally, the data dictionary page describes each metric used by Market Maven, as well as various metadata about the metric, such as source, definition, and aggregation level, to provide the user with complete transparency about this application's data. With this information, users are equipped with knowledge about the source and nature of these data, which may ultimately increase their trust in and comfort with *Market Maven* and lead to more engagement with the tool.

For accessibility, a common header on each page grounds users and allows for effective and simple navigation between areas of the site, allowing the user to access any parts of the tool from whatever page they are currently on.

3.3 Architecture

There were several technologies used to develop *Market Maven.* In this section, we will describe the technologies that were used, as well as discuss the data and algorithms that were collected and implemented, respectively.

The datasets used in the creation of *Market Maven* were extensive. The amount of metrics and records available within the Census API are incredibly large and a great deal of time was spent exploring which variables were relevant for this application. Keeping the end user in mind, the decision was ultimately made to only retrieve and store the metrics that were relevant to the visualizations and information we wanted to present in the tool. For *Market Maven* to be successful, the application needed to be focused and responsive.

There were also steps taken to reduce the amount of data based on the website limitations. For example, the migration dataset contained over 500K edges (1 edge for each county-county migration link), Instead of visualizing all of these edges, we decided to summarize this information by county in our Population Ins/Outs/Net Movement metrics. Additionally, we had the intention of working with a network graph of almost 10M edges, 1 edge for each county-to-county interaction. The decision was made, however, to visualize these data points on a scatter plot by their first 2 principal components, providing an alternative visualization than the originally planned network graph. We believe this decision was necessary to keep the website responsive for our users.

As data is foundational to every analytics project, our approach was to collect as much information up front before the build of the models and interface. A large multitude of data sources were considered, and after our review, we determined that based on the information available, a common granularity of 'county', worked well for this project. The data sources are pulled from a combination of multiple Census API calls, to CSV files collected from reputable organizations. A varying degree of data cleansing and wrangling needed to occur as well, which was performed using the *Pandas* library in Python. Finally, the data was written into a *SQLite* database for table storage and the creation of views.

Python was the primary tool used to develop *Market Maven.* For the front-end, we used the *Dash* library to build the graphical user interface and interactions, and serve *Market Maven* as a web application in the browser. The choropleth map of the U.S. shown on the home page was implemented using *Dash-Leaflet*, which is a Python wrapper around the *Leaflet* JavaScript library. We selected *Leaflet* to implement this feature because it allowed for our application to respond to user interaction events (e.g., mouse clicks, hovers, etc.) on the map. This functionality was necessary for the application to intelligently respond when a user clicks on the map to select a county, for example. The clustering visualization that uses a trained k-means model to display counties similar to the one the user has selected was implemented using the *Plotly* visualization library. While *Leaflet* excels at providing interactive maps, *Plotly* was a more suitable technology for implementing the clustering visualization because of its deep customization options and ability to integrate easily with models trained using the *Scikit-Learn* library.

In addition to providing the functionality for Market Maven's client-side front-end, *Dash* was also used to implement the application's back-end server infrastructure, in combination with a *SQLite* database to store the application's data. All manipulation of the data, whether to prepare the data storage in the *SQLite* database or to alter it in response to users' real-time interaction with the application, was handled by the *Pandas* and *NumPy* libraries.

Finally, the clustering model used to identify similar counties and to make recommendations was implemented using the *Scikit-Learn* library. Specifically, the KMeans method was used to generate the clusters and KDTree was used to query the k nearest neighbors. Missing data was imputed using the weighted average of the 10 nearest neighbors (i.e. counties) to the county selected by the user. The closer the neighbor, the larger the weight applied. This weighting scheme was implemented using the *KNNImputer* method provided by the *Scikit-Learn* library. Extreme outliers, including data points reporting z-scores greater than 50 (i.e,. observations 50 standard deviations away from the mean observation), were removed before modeling the data. These extreme outliers were few and far between and an investigation suggested that the few that did exist were likely due to data entry errors. Once all outliers were removed, the data were scaled using the *StandardScaler* method from the *Scikit-Learn* library. This method converts each observed data point to a z-score. The importance of scaling the data before modeling cannot be overstated because scaling mitigated the impact that features with observations larger in magnitude exhibited on the model relative to features with observations smaller in magnitude. Counties are then clustered into 60 groups which we arrived at using an elbow graph to determine an appropriate inflection point, where adding more groups led to diminishing returns in lowering the sum of squared differences of the points from their cluster centers.

The costs to collect data, create the tool, and host the tool are minimal. Most items have been free of cost including the sourcing of data from APIs, downloading data from other free sources, and using open source programs and libraries. The only incurred cost is the cost to host the website at PythonAnywhere, which is minimal ($5 a month).

4 Experiments and Evaluation

Our first experiment took the form of a user study. This user study was primarily designed to determine whether there is a difference between groups that interacted with *Market Maven* first and then were asked their opinions on the current state of real estate market research or vice versa. The 49 participants were randomly divided into two groups comprised of 24 members in group A and 25 members in group B. Group A was asked to interact with *Market Maven* and answer a few questions about it first. Then they were asked to explore existing tools such as *Zillow* before answering a few questions about existing market research tools. Group B was asked to follow these steps in reverse. Both the general survey and the *Market Maven* survey were comprised almost entirely of questions that asked the participant to give their response on a 1-5 scale, 1 being strongly disagree and 5 being strongly agree. There were additional, open ended and optional questions that asked for names, comments and/or feedback. Since the two groups were asked to complete tasks in opposite orders they served as a sort of control for each other.

Once all the data had been collected, F and t tests were employed to determine whether there were statistically significant differences between the two groups. F tests were applied to each group's answers as a whole as well as to different groups of questions that focused on specific topics such as "innovativeness", usefulness and more. If the F statistic was greater than the F critical one-tail value, we concluded that the answers provided by the two groups for the same questions had unequal variance and conducted a t test assuming unequal variance. If the F statistic was less than the F critical one-tail value, we accepted the null hypothesis that the answers between the two groups had equal variance and t tests were conducted once using an equal variance assumption and again with an unequal variance assumption. These steps were repeated for the Market Maven survey and all of its question categories as well, some of the topics including likelihood to adopt, usefulness of our map feature and usefulness of our comparison feature to name a few. In general the F statistic for both groups' answers to the general market research survey was 1.10. The F critical one-tail value was 2.10. Both t tests using equal and unequal variance assumptions gave two-tail p-values of 0.38. For the responses to the *Market Maven* survey overall, the F statistic was 0.76 and the F critical one-tail value was 0.49. The two-tail p-value of the t test under the unequal variance assumption was 0.92. The outcome of all of these tests strongly suggested that there was no statistically significant difference between the two groups. In every t test conducted the two-tail p-value was greater than the significance level, 0.05 and the t statistic fell between the t critical two-tail values confirming our results. Therefore we accepted the null hypothesis that there was no difference between the groups' means. In other words, whether a group was shown *Market Maven* first or asked about their opinion on the status quo of real estate market research first made no difference in how each pool responded to both surveys. This meant that we were able to combine and treat both groups as one when aggregating statistics.

Our next experiment involved analyzing the results collected from all 49 participants. The mean, median and coefficient of variation of all responses to each of the questions and question categories were aggregated.

The median gave us an idea of where answers fell on our 1-5 scale and the mean gave an indication of whether the responses leaned towards an answer if caught in between ratings. The CV gave an indication of whether the responses were more or less in agreement or varied widely, which for the most part the answers did not seem to do. In general, participants were neutral when it came to the sufficiency of current resources with a slight lean towards disagree when asked specifically on the availability of broad market information versus specific properties/listings. They agreed with a slight lean towards strongly agree that *Market Maven* would be useful to real estate professionals as well as the general population. However, participants were neutral on replacing other tools completely with *Market Maven* or using *Market Maven* as their sole resource. In general, the participants agreed that current tools are user friendly and interesting. However, there was a slight lean towards neutrality on the novelty and visual appeal of current resources. In contrast, participants agreed, but not strongly, that the prototype of Market Maven was visually interesting and easy to use. Participants agreed that current resources are affordable or free but were neutral on the convenience of the tools. In both cases, the responses leaned towards neutral. In terms of *Market Maven's* specific features, the majority strongly agreed or agreed that the map feature was useful and interesting. In general, responses agreed that the details page was interesting and useful and strongly agreed that the comparison feature was useful. Perhaps the most contentious feature was the cluster feature. Due to the fact that we clustered our counties based on a novel combination of data sets with no existing labels, user feedback was critical in determining whether our clusters provided value. Participants agreed that the cluster feature was innovative, interesting, and provided useful recommendations. However, there was room for improvement. While the majority thought the recommendations were useful, this question had one of the highest CV values at around 22.93%, meaning that there was a somewhat split opinion. The answers in general to the *Market Maven* specific questions were less spread out than the answers to the general market research survey. The general survey had an average CV of around 24.22% while the *Market Maven* results had an average CV of around 18.26% meaning subjects were more aligned in their responses on Market Maven as a whole. Overall our pool of participants agreed, with a slight lean towards strongly agree, that they would recommend *Market Maven* to others.

Our final experiment focused on the evaluation of the data used to build our k-means model, which we used 2 principal components for in order to plot in 2 dimensions. Several questions we hoped to answer with this experiment were: Should we have thought of a different way to visualize the data? How was the quality of the features we chose? Could the data be used to develop a more methodical approach to improving the real estate market research experience and improve the feedback received from users? Now that we had an indication, based on our participant study, that our users largely agreed that the recommended counties generated by our clustering algorithm were useful, which of the features contributed the most to useful recommendations and which features did not?

In order to attempt to answer some of these questions, further analysis was conducted on the principal components fit to our imputed, outlier-removed, and scaled data. Taking a look at the explained variance ratio of our 32 principal components, the first 18 explained around 99% of variance while the first 2 only explained a little over 40%. This told us we might have benefited from looking for alternate methods for visualizing the data. Perhaps we could have plotted each county as a node and connected them using similarity scores for edge length, and by only including the edge with the highest similarity score for each node. However, this result also told us that we chose a good mix of uncorrelated and informative features by grouping data into categories based on our literature survey and making sure we had roughly even amounts of data in each category.

The first 18 components became the focus of the following steps. The components and their coefficients were transposed, with the new index being each of the 32 features of our dataset. For each feature the sum of the absolute value of its coefficients was aggregated across all 18 components, weighted by the explained variance ratio of each component. The absolute value was taken because the magnitude is irrespective of direction. The final results were stored in a dataframe and sorted from largest to smallest. This revealed that features such as educational attainment of bachelors or higher, percent of population commuting using eco-friendly methods and density of urban populations were among the more informative features. Features such as percent of population unemployed, percent of population in poverty and per capita income had the lowest coefficients but it did not mean they were less important. The results made sense and were surprising. It seems that for our population, large, eco-friendly and well educated counties were areas of interest. Lower values for unemployment and poverty made sense as those factors are seen as negative in our society when high, but it was surprising that per capita income was not as big of factor when it came to identifying similar markets. We could use these insights to fine tune our model.

Our experiments worked in conjunction with each other to answer questions about how *Market Maven* could be improved, where *Market Maven* succeeded, and whether the clustering model provided value and how it could be enhanced.

5 Discussion

There were many features we would have liked to have included but were unable to implement in the application. We would have liked to include additional metrics, especially cost of living and real estate volume and cost. However, some of these metrics were either unavailable or difficult to acquire within the project timeline. We would have liked to have displayed more information on the details page, such as trends by metric and category scores. Finally, an important feature we had to cut was the personalization of the clustering analysis. Here, a user would select the attributes that are most important to them using a sliding importance scale for each metric. Then the clustering algorithm would rerun and provide the user with new clusters and a new view of the similarity graph. It is unfortunate that we were not able to include this feature, as this would have added a layer of personalization that would have increased user engagement and likely satisfaction, but it is important to note that for a prototype, this feature would have added a large amount of complexity to the tool, as the clustering script would need to rerun on demand. That being said, we are very happy with the decisions made throughout the process and are proud of the features that we ended up including. We would plan to add these features as future enhancements if the team decides to continue development.

As mentioned previously, our experiments revealed that it may have been beneficial to find an alternative method of visualizing the output of our clustering model. In order to plot our data in two dimensions, we used principal components analysis to find the first two components. The results of our experiments on the components suggested that by doing this we may not have taken full advantage of our data. Though users found the output useful and interesting, there was room for improvement.

Data availability was also a challenge because we needed to agree upon a standard level of aggregation for our metrics. There are drawbacks with using counties, such as the fact that counties are not homogeneous and that county definitions differ by state. Additionally, many cities span across several counties, and many counties contain a number of varying real estate markets, further complicating data reporting. However, collecting data at a more granular level proved impossible based on the type and variety of metrics we aimed to include in the tool.

Our literature survey gave us an overview of the real estate landscape, various factors affecting the U.S. markets, and an understanding of the currently available tools in the marketplace. With this, we were able to confirm gaps in the current resource ecosystem and position where *Market Maven* fit and fulfilled customer demand. We were able to improve upon existing applications after understanding their limitations. The literature guided the creation of our problem statement and helped us choose which metrics to include and how to organize them.

6 Conclusion

Market Maven is a novel research tool that fills a niche which is missing among the current tools out there. It allows for a high degree of interaction and information discovery about real estate markets within a single, user-friendly web application. Our map page provides an immersive interface to research and understand data at the market macro level. The details page provides comparison capabilities and additional context for the user to interact with and explore in an engaging way. Finally, our clustering page introduces advanced analytics to users in a creative way, encouraging users to discover new insights they may not have surfaced otherwise.

Our experiments and evaluations have shown that that there is room for improvement, particularly within our clustering algorithm and accompanying interface but also that we succeeded in choosing an informative and diverse set of metrics. We also learned that our audience is not certain whether the resources available are enough for them to effectively research real estate markets but they agree that *Market Maven* would be a welcomed addition to their toolbox.

6.1 Distribution of Effort

Keith, Dan, and Max have contributed equally to all aspects of this project. Ruoyi only contributed to the project's early ideation, COVID-19 section in our literature review, and early template of the data dictionary page. All of these contributions had to be redone by other members of the group for lacking effort.

References

[1] Florida, Richard. "Bohemia and economic geography." Journal of economic geography 2, no. 1 (2002): 55-71.

[2] Glaeser, Edward, and Joseph Gyourko. "The economic implications of housing supply." Journal of Economic Perspectives 32, no. 1 (2018): 3-30.

[3] Miller, Norm G. "Workplace trends in office space: implications for future office demand." Journal of Corporate Real Estate (2014).

[4] Yoruk, Baris. "Early Effects of the COVID-19 Pandemic on Housing Market in the United States." Available at SSRN 3607265 (2020).

[5] Home, O. E. C. D. "Housing Amid Covid-19: Policy Responses and Challenges."

[6] D'Lima, Walter, Luis A. Lopez, and Archana Pradhan. "Covid-19 and housing market effects: Evidence from us shutdown orders." Available at SSRN 3647252 (2020).

[7] Fu, Yanjie, and Hui Xiong. "A discovery system for finding high-value homes." In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1612-1615. IEEE, 2015.

[8] Sun, GuoDao, RongHua Liang, FuLi Wu, and HuaMin Qu. "A Web-based visual analytics system for real estate data." Science China Information Sciences 56, no. 5 (2013): 1-13.

[9] Li, Mingzhao, Zhifeng Bao, Timos Sellis, Shi Yan, and Rui Zhang. "HomeSeeker: A visual analytics system of real estate data." Journal of Visual Languages Computing 45 (2018): 1-16.

[10] Nelson, Garrett Dash, and Alasdair Rae. "An economic geography of the United States: From commutes to megaregions." PloS one 11, no. 11 (2016): e0166083.

[11] Johnson, Kenneth M., and Richelle L. Winkler. "Migration signatures across the decades: Net migration by age in US counties, 1950-2010." Demographic Research 32 (2015): 1065.

[12] McNamara, Dylan E., Sathya Gopalakrishnan, Martin D. Smith, and A. Brad Murray. "Climate adaptation and policy-induced inflation of coastal property value." PloS one 10, no. 3 (2015): e0121278.