Utah Single Family Residence Median Sales and Future Median Sales Projections

JDee Warren - 001463291

Western Governors University

Table of Contents

A. Project Highlights

A real estate firm (fictitious) requested data that examines past single-family median sales versus future values. The question they would like addressed is, based on Zillow's last ten years of housing sales data, will the median sales price for single-family residences in Utah for the five largest cities increase over the next twelve months?

The project's scope will be using Jupyter Notebooks, which uses Python coding to analyze historical median single-family home sales provided by Zillow. The data will be cleaned, transformed, and broken down into months and years to decrease the size of the dataset to a manageable size. Once cleaned, create graphs, charts, and a small dataset to illustrate the projected data for the real estate firm that they may use for their marketing campaigns.

Areas that will not be covered in the scope will be data surrounding interest rates or economic changes. Due to the infrequent changes in interest rates and economic changes, it would be difficult to include those items due to their uncertainty.

The request was accomplished using Python coding in Jupyter Notebooks to complete the analysis. Python offers a free programming platform with access to libraries that range from statistical analysis to data visualization and machine learning, to name just a few. The data provided by Zillow.com provides sixteen years of single-family median housing data for most regions (cities) in the United States. Using Python libraries such as pandas, matplotlib, numpy, scipy, and stats models, we were able to ingest, clean, analyze, and visualize the data requested by the real estate firm.

Using predictive analytics, we were able to forecast future median sales for the largest regions (cities) in Utah. Using a null hypothesis test, we confirm our findings and solidify the data-driven results for the real estate firm.

B. Project Execution

Project Plan

The project's goal was to provide visuals and data to the real estate firm so they could make an educated decision on marketing for the upcoming twelve months. There were three outlined objectives for the goal.

Objective one was to clean Zillow's housing data. The real estate firm wants to analyze Utah's single-family sales for the five most significant regions (cities). This was accomplished by removing all other states and regions from Zillow's data. We then verify that no remaining null information was included in the data, which could distort the data output.

Objective two was to provide visual future sales data to the real estate firm regularly and consistently as data becomes available. This was accomplished using a linear graph, bar chart, and data source indicating the historical and future projections. Delivery will be set to monthly as the data is only updated once a month.

Objective three was to perform a null hypothesis test to confirm our future projection results. The null hypothesis test examined a two-sample set of two separate periods of time (2014-2016 and 2022-2024). The first null test established that there were no value increases during given times. The second test examined for significant increases during the same given times. They provided a t-test, p-value, and alpha values. If the p-value is less than the α = alpha value of 0.05, the null hypothesis is rejected and supports the initial hypothesis.

All three objectives were accomplished, and there were no variations from the original project objectives. The first objective was to clean the data provided by Zillow. As Zillow regularly updates and verifies their data, very few alterations are necessary. The data was drilled down to the five specific regions (cities) in Utah, and all nulls were removed. A linear graph, bar

chart, and data source were created for the second objective. And for the third objective, a null hypothesis was completed and confirmed that the outcomes rejected the null hypothesis.

Project Planning Methodology

We followed the Waterfall methodology as expressed. The Waterfall methodology uses six steps: requirements, design, implementation, testing, delivery, and maintenance.

- For the requirements, we met with stakeholders and defined their requirements for the data. We gathered the data from Zillow.com, which included over 16 years of median single-family sales.
- The design focused on detailing how we would gather, clean, and select pertinent fields, sort the data necessary for the request from the Zillow dataset, and design how the graphs would appear.
- Next comes implementation, where we would use Python to perform predictive analytics on cleaned and transformed data and generate linear and bar charts along with a data source on future trends.
- Once the graphs, charts, and data sources were completed, we performed a null hypothesis test to determine the validity of our outcomes.
- As testing was completed, we could proceed to the delivery phase. As the information from Zillow is updated monthly, a monthly delivery is recommended. This also required sign-offs from the stakeholders.
- The final requirement of the waterfall method is maintenance. We recommend regular documentation and versioning control to allow for future changes.

No changes or variations from the original waterfall methodology were needed to complete the desired outcome.

Project Timeline and Milestones

There were no deviations from the original timeline and milestones provided. Weekends were within the anticipated completion dates, and no additional time was required to complete the project. The chart below illustrates the project's projected, anticipated, and completed dates.

Milestone	Duration (hours or days)	Projected start date	Anticipated end date	Completed date
Project Selection / Requirements	3 Days	*08/04/2024*	*08/07/2024*	*08/07/2024*
Project Approved / Design	1 Day	*08/07/2024*	*08/08/2024*	*08/08/2024*
Zillow Data Download	1 Hour	*08/08/2024*	*08/08/2024*	*08/08/2024*
Implementation • Clean • Transform • Predictive Analytics • Graphs • Data	3 Days	*08/09/2024*	*08/14/2024*	*08/14/2024*
Testing	1 Day	*08/14/2024*	*08/16/2024*	*08/16/2024*
Delivery	1 Day	*08/16/2024*	*08/16/2024*	*08/16/2024*
Maintenance	1 Day	*08/19/2024*	*Ongoing/Quarterly*	*Ongoing/Quarterly*

C. Data Collection Process

The data was collected by filtering the specific criteria: median sales price (Raw, SFR only, Monthly) from https://www.zillow.com/research/data/. The data is downloaded as a .csv file. There were no changes or variations to the data collection from the original project plan.

The data provided by Zillow is a quality product and provides the necessary fields to complete a thorough analysis of the housing data. While some regions (cities) are missing median sales value, they appear absent in the earliest years, which could be due to how recently

those regions made the data available. The data used in this project is complete and readily available. According to the research principles provided on their About page they state, “Zillow Research benchmarks findings against outside datasets whenever possible to ensure accuracy and appropriate context.” (Zillow, n.d)

The data and information used for this project are free and open source to the public. There are no data governance, privacy, security, ethical, legal, or regulatory compliance considerations required for the data being used or stored.

C.1 Advantages and Limitations of Data Set

One of the advantages of the data set for this project was using Zillow’s compiled information. Zillow is a national company that has gathered public data from most major cities nationwide and compiled it into a single source. In addition to gathering data, they have compiled more than sixteen years of housing sales data and continually add to it monthly, making the dataset span multiple years for a comprehensive analysis. For this specific project, we were able to look at median sales values for the five largest regions in Utah, dating all the way back to 2014.

One disadvantage of the data set was its limited information outside the median sales price. While there are over sixteen years and hundreds of regions of data compiled into one data set, there are no other metrics provided. While this project did not account for significant interest rates or economic changes, no other data was provided to indicate why sales prices were increasing. This may have indicated why there may have been fluctuations in the historical information.

D. Data Extraction and Preparation

To access the data, we navigate to https://www.zillow.com/research/data/ under the Sales heading and set criteria for median sales price (Raw, SFR only, Monthly) from the drop-down menu. This will generate a .csv file that can be saved/exported to a local drive. Once the file has been saved, it can be read into a Python environment to prepare it for use. Once read into Python, the file will be trimmed down to include only the five largest regions (cities) in Utah. A null value check will then be performed, and null values will be removed. The data is then transposed so it can be used for the charts. As the file contains information for the past 16 years, we removed any dates older than ten years. Finally, the date column is updated to a datetime field.

E. Data Analysis Process

E.1 Data Analysis Methods

With the data prepared and cleaned, we employed predictive analysis to determine the future Utah median single-family sales for the five largest regions (cities). Using the ten years of historical data provided by Zillow, we applied forecast fit to plot a projection of the next twelve months, which is included in the graphs for representation. This analysis did not take into account any economic or interest rate changes that could occur during the projected outline due to the unpredictability and frequency of those changes.

The linear graph represents the historical and forecasted data the real estate company requested. It displays a ten-year growth pattern followed by twelve months of increasing values represented as a trendline. The visualization makes it easy for any stakeholders to make a quick conclusion of the upcoming twelve months. Based on the forecasted trendlines, the median single-family sales prices will continue to increase.

To provide a deeper look at the upcoming twelve months, we applied the same forecasting method. We created a bar graph showing each region's (cities) twelve-month forecast value by month. The bar graph easily indicates an upward trajectory of each region's (cities) median values. The bar chart makes it easy for any stakeholders to make a quick visual analysis of the upcoming twelve months. The graph and chart support our hypothesis that median single-family sales prices will likely continue to increase over the next twelve months, barring any interest rate or economic changes.

In addition to the bar graph, we used the twelve months of forecasted data to create a data source. The data source, which can be made available to stakeholders, contains the increasing numerical values for each state by month. This can also be a valuable tool for the stakeholders to use in their decisions and upcoming marketing campaigns.

To validate our original hypothesis, we employed a null hypothesis test to determine if there were any differences if placed under a different scenario. With a tremendous amount of data spread over ten years, we completed a two-sample test with smaller and separate time frames in our hypothesis. Our first hypothesis stated that the values have not increased during the 2014-2016 time frame or the 2022-2024 time frame. An alternative hypothesis looked at the same time frames but looked for significant differences in the sales prices. If the p-value was less than the α = alpha value of 0.05, the null hypothesis would be rejected and support the initial hypothesis that the median single-family home sales will increase over the next twelve months.

E.2 Advantages and Limitations of Tools and Techniques

For this project we used Python coding in Jupyter Notebooks to complete the analysis. Python offers a free programming platform with access to libraries that range from statistical analysis to data visualization and machine learning, to name just a few.

One of the major advantages of using Python coding in Jupyter Notebooks is the large set of libraries from which you can draw. For this project, we employed libraries such as pandas, matplotlib, numpy, scipy, and stats models, allowing us to use statistical equations and build graphs and charts rather than calculating them in long form.

A limitation to using Python in Jupyter Notebook is how slow it runs code. I noticed several times while writing the code and testing that it would take longer to run iterations. While the delays were evident, they were not harmful to the project. As we are working with only one data source, there should be no issues providing results for the real estate firm. If this were a more extensive project with multiple data sources, longer delays could impact delivery.

E.3 Application of Analytical Methods

The analytical method for this project will use predictive analysis. Catherine Cote of Harvard Business School online provided this description of predictive analysis. “Predictive analytics uses data to predict future trends and events. It uses historical data to forecast potential scenarios that can help drive strategic decisions”. (Cote, 2024)

We analyzed ten years of median single-family home sales in this scenario to determine how the next twelve months will trend. With predictive analytics, Python uses forecasting to help us determine future trends for the housing market in Utah. We created a model to forecast twelve months of data using the ten-year historical data. The information was verified through the generated linear chart, accompanying bar chart, and data source.

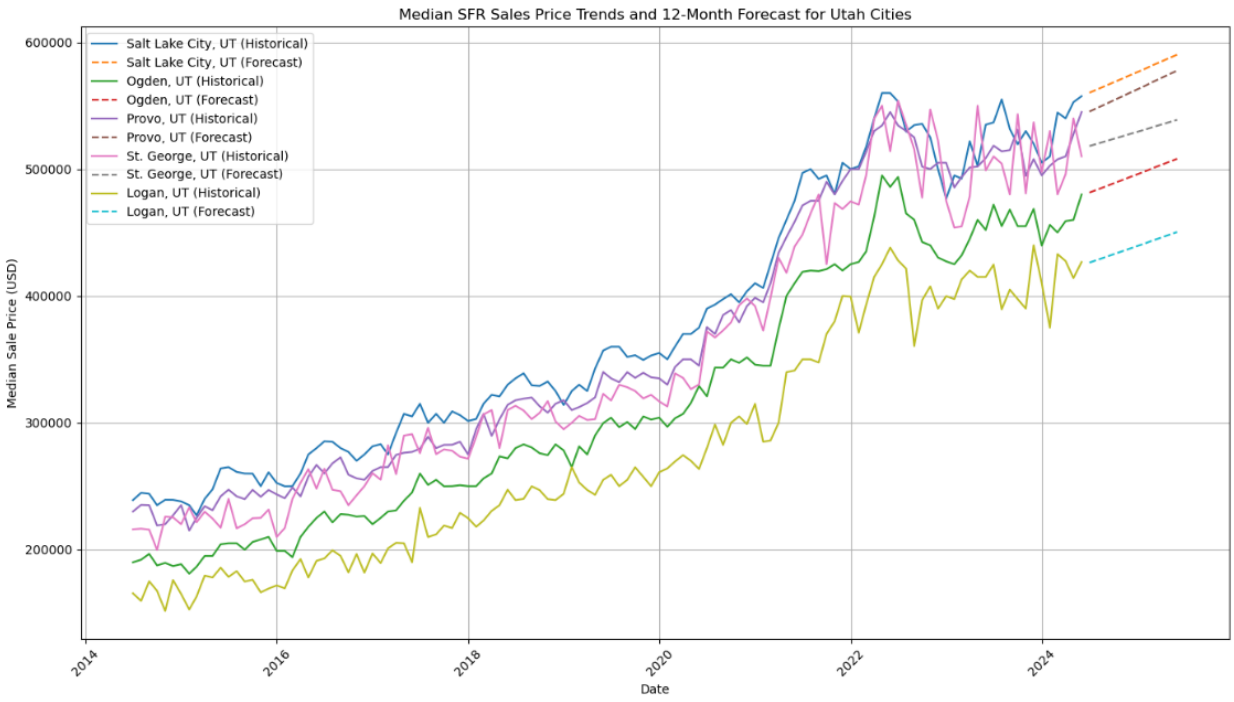
F Data Analysis Results

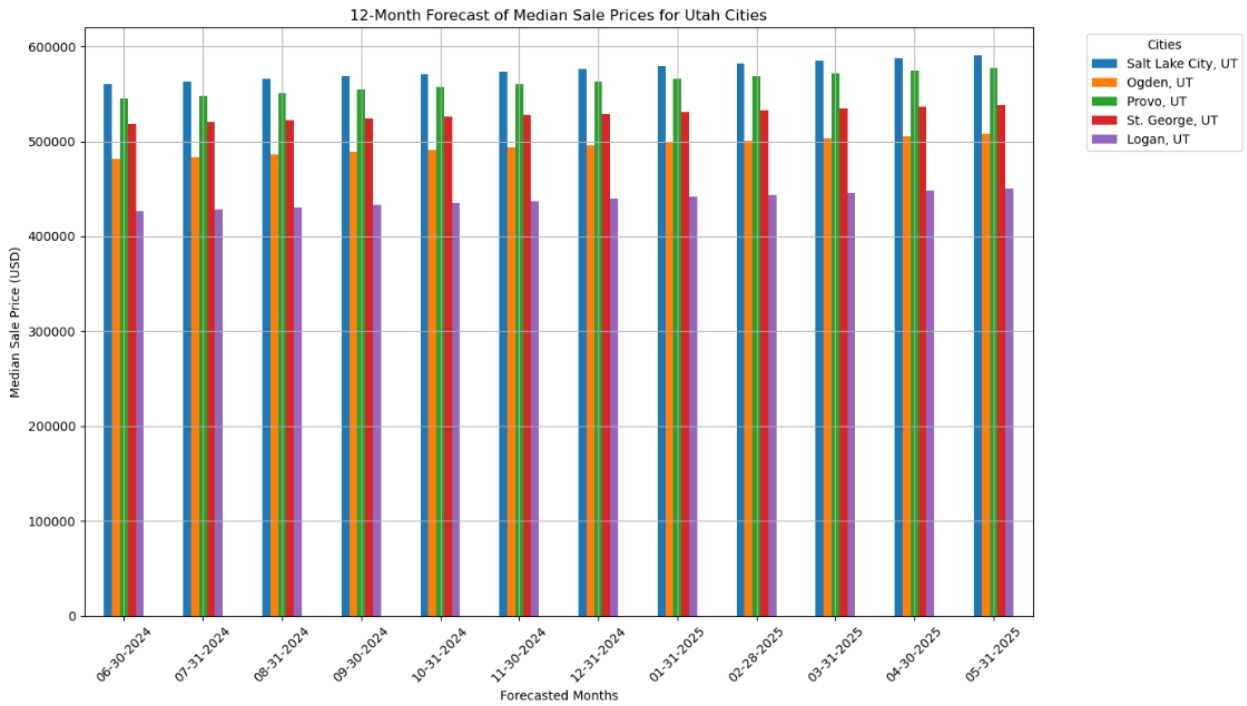
F.1 Statistical Significance

The methods and metrics used to analyze the Utah housing market include a linear chart, bar graph, and hypothesis testing using t-stat, p-value, and alpha value.

The linear and bar graphs represent the ten-year median sales data for single-family residences. Additionally, using the regression model, we can apply forecast fit and plot a projection of the next twelve months, which is represented as a trendline. This analysis does not take into account any economic or interest rate changes that could occur during the projected outline due to the unpredictability and frequency of those changes.

Our linear and bar graphs below, indicate that median sales prices for single-family residences have increased over the last 10 years and will continue to increase for twelve months. The historical data, along with the trendlines, show growth for each of the cities selected.





	Salt Lake City, UT	Ogden, UT	Provo, UT	St. George, UT	Logan, UT
2024-06-30	560216	481484	545363	518192	426280
2024-07-31	562942	483903	548298	520067	428470
2024-08-31	565669	486323	551233	521942	430660
2024-09-30	568395	488743	554168	523817	432850
2024-10-31	571122	491162	557104	525692	435039
2024-11-30	573849	493582	560039	527566	437229
2024-12-31	576575	496002	562974	529441	439419
2025-01-31	579302	498421	565909	531316	441609
2025-02-28	582029	500841	568845	533191	443799

For the statistical test on the housing data, we employed a null hypothesis test to determine if there are any differences if placed under a different scenario. Our first hypothesis stated that the values have not increased during the 2014-2016 time frame and the 2022-2024 time frame. An alternative hypothesis will look at the same time frames but indicate that there is a significant difference in the sales prices. If the p-value is less than the α = alpha value of 0.05, the null hypothesis is rejected and supports the initial hypothesis that the median single-family home sales will increase over the next twelve months.

Based on the null hypothesis test that was performed on each city, we can reject the null hypothesis. All p-values were less than the α = alpha value of 0.05, which indicates that the null hypothesis can be rejected and confirms our initial hypothesis that median values will continue to increase in the coming months.

	City	t-stat	p-value	Reject Null Hypothesis
0	Salt Lake City, UT	-52.905435	8.724461e-51	True
1	Ogden, UT	-56.618599	1.836810e-52	True
2	Provo, UT	-67.860906	5.872443e-57	True
3	St. George, UT	-44.475863	1.616742e-46	True
4	Logan, UT	-53.742636	3.572979e-51	True

F.2 Practical Significance

The data, visualizations, and statistics will allow the real estate firm to make educated decisions on their marketing moves. The graphs, charts, and tests all indicate that median sales values will continue to increase over the upcoming twelve months. These findings can be used in marketing campaigns by the real estate firm to persuade buyers to place offers on homes, knowing that median values will continue to increase and add value to their investments. If updated monthly as new data is made available, all tests and graphs provide the information the real estate firm seeks to meet future financial, marketing, and business needs.

F.3 Overall Success

We outlined the project's success based on three items: the ability to collect usable data, clean and transform the data without any errors, and generate functional graphs and charts that can be used for marketing.

For the first item, the data provided is coming from Zillow, which has been a proven source of information in the housing industry. The data remains relevant as it is updated and

appened on a monthly basis with new sales data. As such, the real estate firm can rely on receiving regular monthly updates to the graphs, charts, and data.

The second item, cleaning and transforming the data without errors, relies on accurate coding. This translates to accurately removing nulls, outliers, and inconsistencies in the data. Along with accuracy, consistency should be a form of success. As the data from Zillow is clean, very few items required cleaning. All nulls were removed, the date fields were updated to datetime, and the values were set to integers rather than float. These changes have all been accomplished using Python code.

And lastly, generating functional graphs and charts that the real estate company can use in its marketing campaign. This was accomplished with the generated charts and graphs. The graphs and charts clearly illustrate a trend for the upcoming twelve months and they can insert them into their marketing campaigns moving forward.

G. Conclusion

G.1 Summary of Conclusions

The original question the real estate firm was looking to address was, based on Zillow's last ten years of housing sales data, will the median sales price for single-family residences in Utah for the five largest cities increase over the next twelve months? Based on the output from the graphs, charts, data sources, and testing, all indicate and support the question on if the median sales prices will continue to increase over the next twelve months. While there are interest rates and economic shifts that could skew these numbers, they are generally infrequent and have not been included in this project. The real estate firm should feel comfortable with the findings and be reassured by the regular monthly analysis of the data provided by Zillow.

G.2 Effective Storytelling

Python coding in Jupyter Notebook was an effective tool while compiling the information requested by the real estate firm. It contains the necessary libraries that assist in analyzing the data from Zillow and provides the necessary visuals for their purpose of marketing. From those libraries, we are able to complete predictive analysis, allowing us to look into the future based on the historical information provided.

The graphs and charts clearly and effectively paint a picture to its audiences, indicating an upward trajectory of median sales prices. The real estate firm can incorporate these graphs into marketing materials, making it easier for potential buyers to make educated decisions on one of the largest purchases of their lives. The graphs can also show that they could expect a return on their investment over twelve months, depending on the region they purchase a home.

G.3 Recommended Courses of Action

Based on the analysis results, I suggest two actions to take moving forward. First, I recommend that the real estate firm utilize the graphs to market to prospective buyers. The linear graph clearly illustrates that based on historical sales of single-family residences, the sales prices are increasing and will continue to increase. The bar chart can show prospective buyers the possible return on their investment, indicating the possible increase in value after twelve months.

The second recommendation is to use the same reports for stakeholders. While it is also important for their marketing team, stakeholders are also important in deciding on upcoming company strategies. While the current analysis shows that median sales will continue to increase, there is a possibility that interest rates or economic change will shift the analysis

results, leading to median sales decreasing. If that were to happen the stakeholders would then be able to readjust their marketing strategy.

H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7b278b65-02a5-4844-8d66-b1dc00206153

References

Zillow (n.d). *About Zillow Research.* Zillow.com
Retrieved August 18, 2024, from https://www.zillow.com/research/about-us/

Cote, Catherine (2021, October 26). *What is predictive analytics? 5 Examples.* Harvard Business School
retrieved August 13, 2024, from https://online.hbs.edu/blog/post/predictive-analytics