Data Analytics Capstone Topic Approval Form

Student Name: Krista Moik

Student ID: 011332151

Capstone Project Name: Multiple Linear Regression on USA Real Estate dataset

Project Topic: Using multiple linear regression to determine which variables are most significant to the sale price of homes in Florida

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: Can a multiple linear regression model be constructed based on the research dataset to determine which variables are most significant to the sale price of homes in Florida?

Hypotheses:
Null hypothesis - H_0: A predictive MLR model cannot determine which variables are most significant to the sale price of homes in Florida with an r-squared value of 0.5 or less.

Alternate Hypothesis - H_1: A predictive MLR model can determine which variables are most significant to the sale price of homes in Florida with an r-squared value greater than 0.5.

Context: The contribution of this study to the field of Data Analytics and the MSDA program is to create a predictive model which can determine the variables that have the greatest impact on the sale price of homes in Florida. This study will utilize a multiple linear regression model to analyze the significance of independent variables and identify which variables best predict the sale price of homes in Florida. Multiple Linear Regression models provide insights into the relationship between the independent variables and dependent variable and how changes in the independent variables can affect the dependent variable (Christopher, 2021). Zhang (2020) found that multiple linear regression models could be used to determine influential factors for house prices and effectively predict the price of houses. The pandemic caused many changes in the United States and the world beyond. One result of the pandemic is that many workers became remote employees and began migrating to other areas of the country they would not otherwise have had the opportunity to work from. Florida is among the states where more people are moving to and less are leaving with 2022 Census data indicating that 249,064 people moved to Florida (Davis, 2023). Due to the influx of people, the real estate landscape is heavily affected by increased demand. It is important for realtors, investors, developers, and businesses with employees moving to Florida to gain a better understanding of what factors are affecting house sale prices.

Data: The data needed to be collected for the question was accessed from a publicly available dataset on Kaggle. The author of the dataset obtained it by scraping public information from Realtor.com. Data obtained via web scraping is legal as the information is publicly available, the information is primarily factual in nature, and the information was not used to steal market share or create a similar product (Apify, 2024). The dataset contains 4,069,420 rows across 10 columns that provides the status, sale price, number of bedrooms, number of bathrooms, lot size, city, state, zip code, house size, and previously sold date. The web scraping started for homes sold on 04/01/2022 and is updated periodically to the present. The data set has a 0.14% sparsity.

The link to access this data is: https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

A description of the variables in the dataset is below:

Variable	**Data Type**
Status	Qualitative
Bed	Quantitative

Bath	Quantitative
Acre_lot	Quantitative
City	Qualitative
State	Qualitative
Zip_code	Qualitative
House_size	Quantitative
Prev_sold_date	Qualitative
Price	Quantitative

Limitations: The dataset is limited by the fact that it only provides data going back two years.

Delimitations: The dataset will be delimited by limiting the analysis to the state of Florida.

The dataset is adequate for this study as it contains useful information such as home size, location, and the number of bedrooms (Hello Data, n.d.). For this study, price is the dependent variable, and the remaining variables are the independent variables.

Data Gathering: Data containing home sale prices as scraped from Realtor.com and published on Kaggle will be downloaded in CSV format. As the information is public and the data is being used for research purposes rather than to steal any business from Realtor, use of this data is allowed. The dataset includes qualitative and quantitative variables.

Cleaning and preparing the data will include searching for duplicates, missing values, and outliers. The data will be explored and visualized to determine the best way to handle outliers and missing values, and categorical variables will be re-expressed as numeric as needed. The missing values and outliers will be reviewed to determine if they should be imputed, retained, or excluded from the analysis. Imputing or excluding these values can introduce possible bias and incorrect values, whereas retaining them without change can also affect the model accuracy (Tamboli, 2023). LabelEncoder will be used for the re-expression of categorical variables. After reducing the dataset to home sales only in the state of Florida, the dataset is now reduced to 345,881 rows and has a sparsity of approximately 0.06%.

Data Analytics Tools and Techniques: Design of the study: An Ordinary Least Squares multiple linear regression model will be fitted to the dataset. Variance Inflation Factors will be used to identify multicollinearity between the independent variables. Backwards Stepwise Elimination will be used due to remove the least statistically significant ones, leaving only the most significant variables. This method is reasonable as there are not a high number of independent variables and this is a standard and accepted variable selection method (Smith, 2018). Once the final model is created, the effectiveness of the model will be evaluated via r-squared values, adjusted r-squared values, p-values, the f-statistic, and the mean and root mean squared error values (Sampaio, 2023). Additionally, a Q-Q Plot and residual plots will be used to determine the normality and check whether the assumptions of the model are met.

> **Justification of Tools/Techniques:** Python will be the programming language used over SAS and R. While all three programming languages are sufficient for data analysis, SAS usage is declining due to its excessive costs and closed-source licensing. Additionally, Python generally handles larger datasets faster than R, making Python the ideal choice for this project (Geeks for Geeks, n.d.).

Project Outcomes: The project will seek to create a multiple linear regression model to determine which variables are most significant to the sale price of homes in Florida. Support for the alternative hypothesis that the independent variables of housing characteristics can account for more than 50% of the variance in house prices is found in Shetty (2020) where the researchers were able to explain 99.6% of the variation in market value, and determined that access road, location, and the number of floors were the biggest variables resulting in increased house prices, and lot shape and age of the building were the greatest factors in decreasing the price of the home.

Projected Project End Date: May 1, 2024

Sources:

Christopher, Antony. (2021, January 7). Applying Multiple Linear Regression in house price prediction. Retrieved April 8, 2024, from https://medium.com/analytics-vidhya/applying-multiple-linear-regression-in-house-price-prediction-47dacb42942b

Davis, Elliott, Jr. (2023, December 5). These States are Bringing in More Residents than They're Losing. Retrieved April 1, 2024, from https://www.usnews.com/news/best-states/articles/these-states-are-bringing-in-more-residents-than-theyre-losing

Is Web Scraping Legal? (2024). Retrieved April 1, 2024, from https://blog.apify.com/is-web-scraping-legal/

Sampaio, Vitor. (2023, June 2). Understanding Ordinary Least Squares (OLS): The Foundation of Linear Regression. Retrieved April 1, 2024, from https://medium.com/@VitorCSampaio/understanding-ordinary-least-squares-ols-the-foundation-of-linear-regression-1d79bfc3ca35

SAS vs R vs Python. (n.d.). Retrieved April 1, 2024, from https://www.geeksforgeeks.org/sas-vs-r-vs-python/#:~:text=In%20summary%2C%20SAS%20is%20a,need%20a%20general%2Dpurpose%20programming

Shett, Dheeraj, et al. (2020). Multiple regression analysis to predict the value of a residential building and to compare with the conventional method values. Retrieved March 26, 2024, from https://iopscience.iop.org/article/10.1088/1742-6596/1706/1/012118

Smith, Gary. (2018, September 15). Step Away from Stepwise. Retrieved April 2, 2024, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6#:~:text=The%20stepwise%20regression%20method&text=A%20backward%2Delimination%20rule%20starts,the%20equation%20is%20statistically%20significant

Tamboli, Nasima. (2023). Effective Strategies for Handling Missing Values in Data Analysis (Updated 2023). Retrieved April 1, 2024, from https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

USA Real Estate Dataset. (n.d.). Retrieved March 26, 2024, from https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

What is Multiple Linear Regression? (n.d.). Retrieved April 1, 2024, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-multiple-linear-regression/

What is regression analysis and how is it used in real estate? (n.d.). Retrieved April 8, 2024, from https://www.hellodata.ai/help-articles/what-is-regression-analysis-and-how-is-it-used-in-real-estate

Zhang, Qingqi. (2021, October 29). Housing Price Prediction Based on Multiple Linear Regression. Retrieved April 8, 2024, from https://www.hindawi.com/journals/sp/2021/7678931/

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 4/11/2024

Reviewed by:

Comments: Click here to enter text.