Data Analytics Capstone Topic Approval Form

Student Name: Briana Churchill

Student ID: 011009463

Capstone Project Name: Multiple Linear Regression on Salt Lake County Real Estate Dataset.

Project Topic: Predictive Model for Salt Lake County Real Estate Data

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

Research Question: Using the USA real estate dataset provided via Kaggle can a multiple linear regression model be created to predict housing prices in Salt Lake County?

Hypothesis: H0: A predictive MLR model regarding Salt Lake County's housing market can be made from the research dataset with a model accuracy > 70%. H1: A predictive model pertaining to Salt Lake County cannot be constructed from the USA real estate dataset.

Context: As a final contribution to the MSDA program at WGU, this study is to build a predictive model that will estimate house prices in Salt Lake County. Because there are many factors that contribute to the price of a house such as lot size, location, and more, this analysis requires a method that can utilize multiple variables to make predictions. Simple linear regression models the relationship between a single dependent variable and a single independent variable. Since there are multiple independent variables, multiple linear regression will be used to create the predictive model. MLR is a valuable tool for predicting housing prices because it allows modeling of the relationship between the dependent variable and each independent variable (Geeksforgeeks, 2023). Having the data needed to make housing price predictions can help many businesses and professions. However, this study was created to make predictions to optimize marketing and pricing strategies for real estate investors (LinkedIn, 2024), as Utah has a growing population and is a hot market for real estate investment over the past couple of years (GuidingOurGrowth, n.d.)

Data: The data needed to complete this study was collected from a publicly available dataset on Kaggle.com. The dataset consists of 2,226,382 rows of collected data from realtor.com, "the second most visited real estate listing website in the United States as of 2024" (Sakib, 2024). The data covers listings from March 2022 through March 2024.

The available variables provided in the data set are listing broker, status of listing, price, number of bedrooms, number of bathrooms, property size in acres, street, city name, state name, zip code, house area in square feet, and the previously sold date. This data set was retrieved using the following link:

https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data

Field	Type
Broker	Qualitative
Listing Status	Qualitative
Price	Quantitative
Number of bedrooms	Quantitative
Number of bathrooms	Quantitative
Lot Size	Quantitative
Street	Qualitative

City name	Qualitative
State name	Qualitative
Zip code	Qualitative
House size	Quantitative
Previous sold date	Qualitative

Broker information and street addresses were categorically encoded by the uploader to maintain privacy. All variables provided do not include identifiable information. Limitations: The data set does not include county information or the listing date. Additionally, there are several instances of null values. Delimitations: the data set will be reduced to listings in Utah, lowering the number of rows to 14,557. Additional reductions will be executed to utilize only listings in Salt Lake County, based off cities within the county. Finally, any null values will be dropped.

Data Gathering: The data set will be downloaded from the publicly available CSV file from kaggle.com which contains 2,226,382 instances of house listings in the USA on Realtor.com. Once the data is loaded the null values found in the bedroom, bathroom, lot size and house size variables will be removed. Because these variables may have an important impact on the results, it would be best to remove them rather than fill them with a median or similar values. Missing values found in the previous sold date variable will be filled as "not applicable" so that those instances may still be utilized for review. Removing or filling in the null values is necessary because missing data may lead to bias and loss of precision (Hughes, R. et al., 2019) which could impact the efficiency of the model. Many rows will be eliminated for this analysis, but fortunately the data set has millions of instances, so the model will not suffer from these reductions because there are still ample amounts of rows to work with. Python will be used to clean and prepare the data. Preparation includes creating dummy variables for all categorical variables and normalizing the data.

Data Analytics Tools and Techniques: Initial data exploration will be completed by creating univariate visualizations to provide insights regarding the dependent variable and each individual independent variable. Afterwards visualizations of bivariate exploration will be executed to review each independent variable alongside the dependent variable. After exploring and visualizing the variables feature selection will be performed using variance inflation factor (VIF) and backward stepwise elimination.
Eliminating features based on their VIF value reduces occurrences of multicollinearity, which happens when there is high correlation between two or more independent variables in a MLR which can lead to skewed or misleading results (Hayes, 2024). Variables with a VIF value of 5 or greater will be removed one by one starting with the highest value. Completing this step will check for and correct any multicollinearity present in the data (Statology, 2020).
Backward stepwise elimination will assess the variables based on the statistical significance of the p-value. Any variables found to have a p-value of less than or equal to 0.05 will indicate no statistical significance and will be removed. Removing insignificant variables improves the model.

Each package used for this assessment, and its purpose (Churchill, 2023):

- Numpy
  - Allows mathematical equations needed to transform the data
- Pandas
  - Provides a logical structure/data frame
- Matplotlib & Seaborn
  - Creates the visualizations (ie: pie chart, scatter plots, box plots, etc.)
- Scipy statsmodels
  - Used many times for various scripts related to: Statistical models, including multiple regression models
  - Creating the linear regression plots
  - Completing the variance inflation factor (VIF)
- Sklearn
  - Used to normalize the data to better understand the data in the regression models (minmax scaler)

Justification of Tools/Techniques: Python will be used to create the regression model for the study because unlike R, Python has the scikit-learn library which “provides a convenient implementation of multiple linear regression” (Geeksforgeeks, 2023). SAS was not considered due to its proprietary nature (SASvsRvsPython, 2023).

Project Outcomes: The project will pursue the creation of a multiple linear regression model to predict housing prices in Salt Lake County based off data from Realtor.com. The results will be demonstrated using Tableau, a data visualization tool known for its ease of use (LinkedIn, 2024).

Projected Project End Date: April 25, 2024

Sources:

Churchill, Briana. (2023, Aug 22). *Performance Assessment: Predictive Modeling Task 1.* Assignment for MS Data Analytics Course D208. Western Governors University.
Guiding Our Growth: The Future of Housing in Utah. Retrieved April 14, 2024, from https://guidingourgrowth.utah.gov/the-future-of-housing-in-utah/

Hayes, A. (2024, Mar 29). Multicollinearity: Meaning, Examples, and FAQs. Retrieved April 16, 2024, from https://www.investopedia.com/terms/m/multicollinearity.asp

How can predictive analytics help you make profitable real estate investments? (2024, Mar 20). Retrieved April 14, 2024 from https://www.linkedin.com/advice/3/how-can-predictive-analytics-help-you-make-profitable-4wbec#:~:text=Improved%20Decision%20Making%3A%20Predictive%20analytics,investment%20decisions%20in%20real%20estate.&text=Risk%20Mitigation%3A%20By%20analyzing%20historical,associated%20with%20real%20estate%20investments.

How to Calculate VIF in Python. (2022, July 20) Retrieved April 14, 2024, from https://www.statology.org/how-to-calculate-vif-in-python/

Hughes, R., et. Al., (2019, Mar 16). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Retrieved April 14, 2024 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6693809/

ML | Multiple Linear Regression using Python. (2023, Jan 25). Retrieved April 15, 2024, from https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/

Multiple Linear Regression using R to predict housing prices. (2023, Nov 6). Retrieved April 16, 2024, from https://www.geeksforgeeks.org/multiple-linear-regression-using-r-to-predict-housing-prices/

SAS vs R vs Python. (2023, June 12). Retrieved April 14, 2024, from https://www.geeksforgeeks.org/sas-vs-r-vs-python/

Sakib, A. (2024, Mar 30). USA Real Estate Dataset. Retrieved April 13, 2024 from https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

What are the key features and benefits of using Tableau for data visualization? (2024, Mar 20). Retrieved April 16, 2024, from https://www.linkedin.com/advice/0/what-key-features-benefits-using-tableau-data-visualization

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: 4/16/2024

Reviewed by:

Comments: Click here to enter text.