Data Analytics Capstone Topic Approval Form

Student Name: Darrell Friday Jr

Student ID: 011345513

Capstone Project Name: Multiple linear regression on real estate data set.

Project Topic: Predictive model for median house value in California.

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

Research Question: Can a multiple linear regression model be created on the research data set?

Hypothesis: **Null hypothesis-.** A MLR model cannot be constructed on the research data set.
Alternate Hypothesis-. A MLR model can be constructed on the research data set with an accuracy greater than 70%.

Context: The contribution of this study to the field of Data Analytics and the MSDA program is to create a predictive model that can estimate the median housing value in California so real estate developers can accurately appraise their property. This study will utilize a multiple linear regression model to analyze the significance of predictor variables and their correlation to the median housing value. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data (Yale University, 2019). Abdulhafedh (2022) found that multiple linear regression can effectively predict the response variable with big datasets and a large number of predictors. The researchers hypothesized that multiple linear regression can precisely predict house prices with a big dataset and large number of both categorical and numerical predictors.

Data: The data collected for this study is publicly available information provided by the U.S. census bureau (*California Housing Prices*, n.d.). The data contains information from the 1990 California census. The data set contains 20,640 rows.
The data set contains the following variables of longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value, ocean_proximity. The data set is available through kaggle.com.
https://www.kaggle.com/datasets/camnugent/california-housing-prices.
The breakdown of the variables is shown below.

Field	**Type**	**Context**
longitude	continuous	independent
latitude	continuous	independent
housing_median_age	discrete	independent
total_rooms	discrete	independent
total_bedrooms	discrete	independent
population	discrete	independent
households	discrete	independent
median_house_value	continuous	dependent
ocean_proximity	categorical	independent

The limitations of the study are that the data is collected from the 1990 California census and does not include data from more recent years. There are no de-limitations to this study. All independent variables will be studied including all available observations (Abdulhafedh, A. (2022)).

Data Gathering: Duplicate rows will be dropped. Missing values will be identified and imputed. Missing values provide a wrong idea about the data itself, causing ambiguity. For example, calculating an average for a column with half of the information unavailable or set to zero gives the wrong metric (Dancuk, 2021). The overall data sparsity is less than 1%.

Data Analytics Tools and Techniques: The design of the study: 1. A Q-Q plot will be used to determine the normality of the data. 2. Categorical variables will be encoded with dummy variables. 3. A Multiple Linear Regression model will be constructed using all variables. The model will utilize stepwise regression using backward elimination. The process of stepwise regression can begin by selecting statistical measures to evaluate the performance of the model. Common indices used in stepwise regression include the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and modified R-squared. The algorithm starts with the full set of features and iteratively removes the least statistically significant feature from the model. This process continues until no more features can be removed without reducing the model's performance (Khadka, 2023). The goals and expectations of this study are to determine which variables are correlated to housing prices, and predict house prices. The presentation layer will consist of univariate and bivariate graphs, box plots, and model summary output.

Justification of Tools/Techniques: Python will be use for the regression model of the data. Python is the appropriate choice due to it's ability to be used in all steps of the data analysis. Python also reduces the time needed to create a working model because it is an interpreted language and has no compile time (Python vs R: Know the Difference, 2021). Kumar (2023) suggests that Python has a stronger set of libraries and tools than SAS.

Project Outcomes: The projected outcome will be a reusable multiple linear regression model for predicting the median house value based on census data of the house and the area it is located. Support for the alternate hypothesis can be found in (Zhang, 2021).

Projected Project End Date: 5/30/2024

Sources:

Abdulhafedh, A. (2022). Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables. *OALib*, *09*(01), 1–21.

Retrieved May 20, 2024, from https://www.scirp.org/journal/paperinformation?paperid=115003

California Housing Prices. (n.d.). Www.kaggle.com.

Retrieved May 20, 2024, from https://www.kaggle.com/datasets/camnugent/california-housing-prices

Dancuk, M. (2021, July 1). *Handling Missing Data in Python: Causes and Solutions*. Knowledge Base by PhoenixNAP.

Retrieved May 20, 2024, from https://phoenixnap.com/kb/handling-missing-data-in-python

Khadka, N. (2023, October 2). *Stepwise Regression: A Master Guide to Feature Selection – Dataaspirant*.

Retrieved May 20, 2024, from https://dataaspirant.com/stepwise-regression/

Kumar, A. (2023, September 26). *SAS vs. R vs. Python: A Data Science Professional's Perspective* . Medium.

Retrieved May 20, 2024, from https://medium.com/@aman19/sas-vs-r-vs-python-a-data-science-professionals-perspective-34416af1d022

Python Vs R: Know The Difference. (2021, October 10). InterviewBit.

Retrieved May 20, 2024, from https://www.interviewbit.com/blog/python-vs-r/

Yale University. (2019). *Multiple Linear Regression*. Yale.edu.

Retrieved May 20, 2024, from http://www.stat.yale.edu/Courses/1997-98/101/linmult.html

Zhang, Q. (2021, October 29). *Housing price prediction based on multiple linear regression*. Scientific

Retrieved May 20, 2024, from Programming. 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: 5/23/2024

Reviewed by:

Comments: Click here to enter text.