**Data Analytics Capstone Topic Approval Form**

*The guidelines for the Data Analytics Graduate Capstone course require a student to demonstrate the "application of academic and professional abilities developed as a graduate student. The capstone challenges students to integrate skills and knowledge from several program domains into one project." As a result, it is highly recommended that your topic* ***should resolve a current or perceived business problem.*** *As stated, you want to exemplify what you have learned in your MSDA program to showcase your skills. Remember, your research topic exemplifies scholarship and research at the highest level and is significant and helpful to potential employers in identifying your abilities. Your research will always be something you can look back on with pride. Finally, it is recommended for students to use publicly available data sets for transparency and external validity.*

*The purpose of this document is to help you clearly state the research question you will be exploring in your capstone project, your project’s scope, and your timeline in order to ensure that these align with your degree emphasis. Without clearly defining each of these areas, you will not have a complete and realistic overview of your project, and it cannot be accurately assessed whether your project will be acceptable for the purposes of these courses. Of course, if this a project that you have already completed at work or elsewhere, this should be easy to fill in! Many students do use a project they have already completed in the past year or two. In that case, you will write the proposal as if the project had not been done yet, and when you report on your project, you will use your complete after-implementation report. If you have not yet done your project, this document can help make sure the scope is within the acceptable range for this capstone. A course instructor will approve this form before submitting this task for evaluation. The task will not be evaluated without a course instructor’s signature. The course instructor may ask for additional information before approving the form.*

*Before submitting this form for approval, please remove all italicized directions in the form.*

***Please only submit a Topic Approval Form that has been signed by a course instructor for evaluation.***

**Student Name:** Darrell Friday Jr

**Student ID: 011345513**

**Capstone Project Name:**  California house price linear model.

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

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

**Research Question:** Can a linear regression model be created based on the census data?

**Hypothesis**: **Null hypothesis**-. A statistically significant predictive model cannot be created from the census data. **Alternate Hypothesis**-. A statistically significant predictive model can be created from the census data.

**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 | continuous | 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.

**Data Gathering:** The data will be downloaded as a publicly available CSV file from the kaggle.com website. Python programming language will be used to clean the data. 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)

**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).

**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) Python has many libraries that are necessary to carry out major data science-related functions. (Kumar, 2023) Python is known for its large and growing ecosystem of libraries and tools. H

ave access to the latest features and techniques ear

**Project Outcomes**: The projected outcome will be a reusable statistically significant multiple linear regression model for 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 (Abdulhafedh, 2022).

**Projected Project End Date**:

**Sources**:

Yale University. (2019). *Multiple Linear Regression*. Yale.edu. <http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm>l

‌

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.

<https://www.scirp.org/journal/paperinformation?paperid=115003>

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

‌ 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.

<https://phoenixnap.com/kb/handling-missing-data-in-python>

‌

Khadka, N. (2023, October 2). *Stepwise Regression: A Master Guide to Feature Selection - Dataaspirant*. https://dataaspirant.com/stepwise-regression/

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

https://www.interviewbit.com/blog/python-vs-r/

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

<https://medium.com/@aman19/sas-vs-r-vs-python-a-data-science-professionals-perspective-34416af1d022>

‌

**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: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.