Data Analytics Capstone Topic Approval Form

Student Name: Elizabeth Sweet

Student ID: 001000431

Capstone Project Name: Linear Regression on D.C. Properties Data

Project Topic: Predictive Model for D.C. Properties Data

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

Research Question: To what extent does Ward and other features affect the price of a home in Washington D.C.?

Hypothesis: H_0: Ward and other features do not statistically significantly affect the price of the house.

H_1: Ward and other features statistically significantly affect the price of the house.

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 house prices in the Washington D.C. area. A regression analysis will be used in the study. Regression analysis is a form of predictive modeling that explores the relationship between a target and predictor variables. (Sunil, 2015) Multiple linear regression will be used to explore the statistical significance of the predictor variables and to what degree they affect the price of the house in Washington D.C. This model can be used by home buyers and investors to compare features and minimize the price paid for property with specific features.

Data: The data used for this study is publicly available through Kaggle.com. The dataset is titled 'D.C. Residential Properties'. The data set contains 158957 real estate records in Washington D.C. (Chrisc, 2018)

The data set includes 49 different variables including number of rooms, number of bathrooms, year built, year remodeled, address, ward, and others. The predictor variables being used in the study are as follows:

Column	Type	Description
BATHRM	Numeric	Number of Bathrooms
HF_BATHRM	Numeric	Number of Half Bathrooms
HEAT	Categorical	Heating Type
AC	Boolean	Air Conditioning
NUM_UNITS	Numeric	Number of Units
ROOMS	Numeric	Number of Rooms
BEDRM	Numeric	Number of Bedrooms
AYB	Numeric	When main portion of building was built
EYB	Numeric	Year of last improvement
STORIES	Numeric	Number of Stories
GBA	Numeric	Gross building area in square feet
STYLE	Categorical	Style Description
STRUCT	Categorical	Structure Description
GRADE	Categorical	Grade Description

CNDTN	Categorical	Condition Description
EXTWALL	Categorical	Exterior Wall Description
ROOF	Categorical	Roof Type Description
INTWALL	Categorical	Interior Wall Description
KITCHENS	Categorical	Number of Kitchens
FIREPLACES	Categorical	Number of Fireplaces
LANDAREA	Categorical	Land Area of Property in square feet
WARD	Categorical	City Ward

Limitations: This study is limited by the accuracy and completeness of the data set. Many of the features have missing information. This data set only includes information for houses and condos. Delimitations: Several features will be removed from this study. Some, like full address, will be removed to protect privacy. Others like sale number, building number, and square will be removed because they are not in the scope of this study. Rows that have the target variable missing, price, will be removed.

Data Gathering: A csv file will be downloaded from Kaggle.com. The data set will be imported into a Jupyter notebook and prepared using the Pandas and NumPy packages.

Data Analytics Tools and Techniques: The study will begin with an exploration of the features to see if they meet the required assumptions of an ordinary least square regression (OLS). A heatmap will be created to inspect the Pearson coefficients between features. A QQ plot will be built to test if the data set is normally distributed. (Samaha, 2020) Next, an ordinary least square model will be built to estimate the relationship between the predictor variables and the target variable (price). (Jr, n.d.)

Justification of Tools/Techniques: Python will be used to create the regression model. Python has the advantage of being easy to use and it offers many libraries to complete any task. (Python Advantages and Disadvantages - Step in the right direction, n.d.) Multiple regression is the appropriate technique to use in this study because the target variable (price) is a continuous numeric variable, and the predictive variables are both numeric and categorical. (Sunil, 2015)

Project Outcomes: This study will create an ordinary least square regression to predict the cost of a house in Washington D.C. given the features included in the model with an acceptable success rate. It is expected that Ward will be the most significantly significant feature.

Projected Project End Date: 5/30/2022

Sources:

Chrisc. (2018). *D.C. Residential Properties*. Retrieved March 2022, from Kaggle: https://www.kaggle.com/datasets/christophercorrea/dc-residential-properties?select=DC_Properties.csv

Jr, D. L. (n.d.). *Ordinary Least Squares Regression*. Retrieved March 2022, from Encyclopedia.com: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/ordinary-least-squares-regression

Python Advantages and Disadvantages - Step in the right direction. (n.d.). Retrieved March 2022, from TechVidvan.com: https://techvidvan.com/tutorials/python-advantages-and-disadvantages/

Samaha, B. (2020, June 14). *My Guide to Understand the Assumptions of Ordinary Least Squares Regressions*. Retrieved March 2022, from Medium: https://medium.com/swlh/my-guide-to-understanding-the-assumptions-of-ordinary-least-squares-regressions-b180f81801a4

Sunil. (2015, August 14). *7 Regression Techniques you should know!* Retrieved March 2022, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

To be filled out by a course mentor:

☒ The research is exempt from an IRB Review.

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

Course Mentor's Approval Status: Approved

Date: 4/21/2022

Reviewed by: Daniel J. Smith, PhD, MBA

Comments: Click here to enter text.