

MSDA Capstone Topic Approval Form

Student Name: Alexander Vaillant

Student ID: 001439197

Capstone Project Name: Linear Regression on LA Airbnb Data

Project Topic: Predicting Airbnb Rental Price using Multiple Linear Regression

Research Question: To what extent do the independent variables of Airbnb rentals predict the rental price in the Los Angeles Market?

Hypothesis:

Null Hypothesis (H_0): A statistically significant model cannot be created to predict the Airbnb rental price.

Alternate Hypothesis (H_1): A statistically significant model can be created to predict the Airbnb rental price.

Context: The contribution of this study to the MSDA program and the Data Analytics field is to create a predictive model which approximates an Airbnb's rental price so that a new host in the Los Angeles market may gauge a potential property's affordability and revenue against competitors. With 32,241 listings in the Los Angeles market, price and the variables with influence on price play a crucial role in the revenue of an Airbnb. In this study, a Multiple Linear Regression model will be utilized to analyze the statistical significance of independent, or predictor, variables which have the most influence on an Airbnb's rental price (dependent variable). When these highly influential predictor variables are known, a host may cater to those areas to attract customers. "Multiple regression allows for a relationship to be modeled between multiple independent variables and a single dependent variable where the independent variables are being used to predict the dependent variable" (Laerd Statistics, 2015). In their paper "Real Estate Value Prediction Using Linear Regression", Ghosalkar and Dhage utilize linear regression to predict the value of real estate (Ghosalkar & Dhage, 2018). Like with real estate value, linear regression can be used to predict the rental price of an Airbnb rental.

Data: The data needed to be collected for this study is publicly available through Inside Airbnb website. "Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world" (Cox, n.d.). The data was compiled from public information on the Airbnb website by Murray Cox. Before any data cleansing and removal is done, the Los Angeles dataset contains 32,241 rows with 74 columns and data sparsity of less than 10%. Host demographic and PII data will be removed before analysis.

This dataset is available through Inside Airbnb here: <http://insideairbnb.com/get-the-data.html>. The study will place a delimitation on the dataset by limiting the neighborhood group to only the city of Los Angeles. There are two other neighborhood groups with observations that will be removed. Another potential delimitation would be the number of features allowed in the model based on stepwise regression. The limitations of this dataset are the presence of host PII and the data sparsity of 10% in both categorical and continuous variables. These limitations can be worked around in the data cleansing process. After PII data is removed, the dataset contains the following 31 usable variables:



WESTERN GOVERNORS UNIVERSITY

Variable	Type	Intention
Id	Continuous	Index
host_response_time	Categorical	Predictor/Independent
host_response_rate	Continuous	Predictor/Independent
host_acceptance_rate	Continuous	Predictor/Independent
host_is_superhost	Categorical	Predictor/Independent
host_listings_count	Continuous	Predictor/Independent
host_has_profile_pic	Categorical	Predictor/Independent
host_identity_verified	Categorical	Predictor/Independent
neighbourhood_group_cleansed	Categorical	Predictor/Independent
room_type	Categorical	Predictor/Independent
accommodates	Continuous	Predictor/Independent
bathrooms_text	Continuous (once cleansed)	Predictor/Independent
bedrooms	Continuous	Predictor/Independent
beds	Continuous	Predictor/Independent
price	Continuous	Response/Dependent
minimum_nights	Continuous	Predictor/Independent
maximum_nights	Continuous	Predictor/Independent
has_availability	Categorical	Predictor/Independent
availability_30	Continuous	Predictor/Independent
availability_60	Continuous	Predictor/Independent
availability_90	Continuous	Predictor/Independent
availability_365	Continuous	Predictor/Independent
number_of_reviews	Continuous	Predictor/Independent
review_scores_rating	Continuous	Predictor/Independent
review_scores_accuracy	Continuous	Predictor/Independent
review_scores_cleanliness	Continuous	Predictor/Independent
review_scores_checkin	Continuous	Predictor/Independent
review_scores_communication	Continuous	Predictor/Independent
review_scores_location	Continuous	Predictor/Independent
review_scores_value	Continuous	Predictor/Independent
instant_bookable	Categorical	Predictor/Independent

The data was compiled from public information on the Airbnb website by Murray Cox. "The data is available under a Creative Commons CC0 1.0 Universal (CC0 1.0) 'Public Domain Dedication' license" (Cox, n.d.). Based on the Creative Commons CC0 1.0 Universal license, a user can "copy modify, distribute and perform the work, even for commercial purposes, all without asking permission" (Creative Commons, n.d.). it can be used for commercial purposes. There are several other Airbnb datasets available on Kaggle with public domain licenses. For example, another dataset can be found at <https://www.kaggle.com/kritikseth/us-airbnb-open-data> with the CC0: Public Domain license. All host PII data will be removed at the start of the analysis to increase privacy.

Data Gathering: The dataset will be downloaded from the Inside Airbnb website in .gz format. From there, the listings.csv.gz can be scraped with a few for loops in Python and the final listings.csv file will be exported to be cleansed of host PII data in excel. The sparsity in this dataset is less than 10%. "Datasets may have missing values, and this can cause problems for many machine learning algorithms" (Brownlee, 2020). In preliminary exploratory data analysis, over half of the 32,241 observations have at least one feature with a missing value. Rather than remove half of the observations, the study will use the KNN



algorithm to impute missing values. In his 2020 paper “kNN Imputation for Missing Values in Machine Learning”, Brownlee states that “the k-nearest neighbor (KNN) algorithm has proven to be generally effective” at predicting and imputing missing values (Brownlee, 2020). This imputation will account for continuous variables. Depending on further analysis in the study, missing values in categorical variables will be replaced with “other” or removed.

Data Analytics Tools and Techniques: Python will be used for the creation, exploration, and evaluation of the Multiple Linear Regression model. To identify the normality of the data, a Shapiro-Wilks test will be used with the SciPy library available in Python. Stepwise regression will be utilized to identify the most impactful variables of the dataset on an Airbnb’s rental price. Before the construction of the model, the cleansed data will be randomly split into training and testing sets of 80% and 20% size, respectively. The training set will be utilized in the model fitting phase while the testing set will be utilized in the model evaluation phase.

Justification of Tools/Techniques: According to a Udacity India Article by Prince Patel, “the main reason for using Python would be readability, versatility and easiness” (Patel, 2018). Since this study is meant to be accessible to new hosts and Multiple Linear Regression is “usually the first machine learning algorithm that every data scientist comes across”, it affords the newer data enthusiasts an easier route to utilize this study in commercial practice (Agarwal, 2018). According to Laerd Statistics, Multiple Linear Regression would be a viable method for this study as it is “used to predict a continuous dependent variable based on multiple independent variables” (Laerd Statistics, 2015).

Application type (select one)

- ☐ Mobile
- ☐ Web
- ☒ Stand-alone

Programming/development language(s) you will use: Python

Operating System(s)/Platform(s) you will use: Windows

Database Management System you will use: Not Applicable

Project Outcomes: The projected outcome for this study is a statistically significant Multiple Linear Regression model that predicts an Airbnb’s rental price based on significant independent variables of an Airbnb. Support for the alternate hypothesis is found in Ghosalkar & Dhage (2018). The Multiple Linear Regression model will be a reduced model consisting of less than 31 variables based on the output of the stepwise regression.

Projected Project End Date: 9/30/2021

Sources:

Agarwal, A. (2018, October 5). *Linear Regression using Python*. Towards Data Science. <https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2#>.

Brownlee, J. (2020, June 24). *kNN Imputation for Missing Values in Machine Learning*. Machine Learning Mastery. <https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/>.

Cox, M. (n.d.). *Get the Data*. Inside Airbnb. <http://insideairbnb.com/behind.html>.



Creative Commons. (n.d.). *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication*. Creative Commons. <https://creativecommons.org/publicdomain/zero/1.0/>.

Ghosalkar, N., & Dhage, S. N. (2018). *Real Estate Value Prediction Using Linear Regression*. Semantic Scholar. <https://www.semanticscholar.org/paper/Real-Estate-Value-Prediction-Using-Linear-Ghosalkar-Dhage/f2308f0a4f0981801b518b9ca2152bcb4c797ad7>.

Laerd Statistics (2015). *Multiple regression using SPSS Statistics*. Laerd Statistics. <https://statistics.laerd.com/premium/spss/mr/multiple-regression-in-spss-20.php>

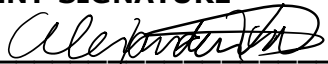
Patel, P. (2018, March 8). *Why Python is the most popular language used for Machine Learning*. Medium. <https://medium.com/@UdacityINDIA/why-use-python-for-machine-learning-e4b0b4457a77>.

INFORM INSTRUCTOR OF:

Potential use of human subjects (Y/N): Yes, but no human subjects are used.

Potential use of proprietary company information (Y/N): Yes, but it is not necessary as the data comes with a CC0 1.0 Public Domain license.

STUDENT SIGNATURE

_____

By signing and submitting this form you acknowledge any cost associated with development and execution of the application will be your (the student) responsibility.

Institutional Review Board Quiz and Approval

Have you read and understood the "Human Subjects FAQ" page and completed the "Human Subjects FAQ Quiz" at the WGU Institutional Review Board (IRB) website?
(<https://irb.wgu.edu/info/Pages/Home.aspx>)

☒ Yes, I have read and understood the "Human Subjects FAQ" and have provided email proof of my completed quiz in appendix A. (<https://irb.wgu.edu/info/Pages/Human-Subjects-FAQ-Quiz.aspx>)

☐ No, I have not completed the Human Subjects FAQ quiz.

Assess whether your capstone proposal complies with WGU's IRB standards for exemption status. Explain why you believe the proposed project complies with the standards for exemption status. If it does not, make arrangements with a course mentor and the IRB for approval.

☒ The research complies with WGU's IRB exemption status because:

- Research involving the collection or study of freely available de-identified existing data
- Research that does not employ methodology on human subjects.

☐ The research requires approval from WGU's IRB because:



WESTERN GOVERNORS UNIVERSITY

☐ Yes, I would like to schedule a conference to discuss my project.

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: 8/16/2021



Reviewed by:

Comments: [Click here to enter text.](#)

