**Data Analytics Capstone Topic Approval Form**

**Student Name:** Mikaela Gardner

**Student ID:** 010939588

**Capstone Project Name:** NYC Airbnb Pricing Linear Regression Analysis

**Project Topic**: The objectives of this data analysis are to identify the key variables that significantly impact the pricing of Airbnb’s in NYC, construct a multiple linear regression model capable of predicting total charges based on relevant factors, and offer insights into Airbnb pricing strategies for competitors while giving recommendations on potential areas for improvement. The analysis will use an [open-source dataset](https://insideairbnb.com/get-the-data) containing data from Airbnb’s in NYC (Inside Airbnb, n.d.).

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

**Research Question:** Can the pricing of Airbnb’s be predicted using a predictive linear regression model based on the research dataset?

**Hypothesis**: **Null hypothesis**: The multiple linear regression model constructed on the research dataset cannot predict the price of Airbnb’s in NYC accurately, with a mean absolute percentage error (MAPE) greater than or equal to 20%. **Alternate Hypothesis**: The multiple linear regression model constructed on the research dataset can predict the price of Airbnb’s in NYC accurately, with a mean absolute percentage error (MAPE) less than 20%.

**Context:** Since its inception, Airbnb has transformed the short-term accommodation market by allowing individuals to trade space, triggering fierce competition from established commercial providers. The emergence of COVID-19 further reshaped the industry, with Airbnb continuing to play a pivotal role. In this evolving landscape, competitors of Airbnb have faced increased challenges, struggling to remain viable amidst the disruptions caused by both Airbnb's presence and the global shut-down (von Briel & Dolnicar, 2021). As the world emerges from the pandemic, both Airbnb and its competitors stand at a pivotal juncture, a critical moment where the business strategies they choose could lead to either prosperity or downfall. By studying recent Airbnb pricing data, competitors can gain insights into market dynamics, identify potential opportunities for pricing strategies, and make informed decisions to adapt to the changing landscape post-pandemic. This study will use Multiple Linear Regression on Airbnb data in order to quantify the impact of each independent variable and predict future prices based on those independent variables. This study will contribute to the field of Data Analytics by creating a predictive model that will give an estimation of NYC Airbnb prices to help competitors adjust their business strategies. Other studies have used Multiple Linear Regression to successfully solve similar problems, such as ‘Multiple linear regression model for predicting bidding price’ which found that Multiple Linear Regression analysis was well-suited for predicting bidding prices based on real-world business data (Petrovski et al., 2015).

**Data:** The data for this project, comprising quarterly data for the last year from Airbnb listings in NYC, was obtained from Inside Airbnb, a platform that provides public domain data on Airbnb listings (<https://1drv.ms/x/c/1b1bf5fe00687082/ETRb9f2qUw1Noj2WOYDdMz4BEOKTi4PDHQDA-8vgFWwkxA?e=yMROby> ). The dataset contains information on various aspects of Airbnb listings, including price, geographical coordinates (latitude and longitude), room type, number of bedrooms, minimum nights required for booking, number of reviews, reviews per month, availability throughout the year, neighborhood group, and review scores rating. Data from this website has been analyzed in many studies, including ‘Impacts of the Peer-To-Peer Market on Tourist Accommodation on the Balearic Islands of Mallorca and Menorca’, which found that there is a negative correlation between the number of Airbnb reviews and degree of hotel occupancy (Benítez-Aurioles, 2020).

The columns to be analyzed:

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Variable Type |
| neighbourhood\_group\_cleansed | Categorical | Independent |
| latitude | Continuous | Independent |
| longitude | Continuous | Independent |
| property\_type | Categorical | Independent |
| room\_type | Categorical | Independent |
| bedrooms | Discrete | Independent |
| price | Continuous | Dependent |
| minimum\_nights | Discrete | Independent |
| maximum\_nights | Discrete | Independent |
| number\_of\_reviews | Discrete | Independent |
| review\_scores\_rating | Categorical | Independent |
| review\_scores\_cleanliness | Categorical | Independent |
| review\_scores\_checkin | Categorical | Independent |
| review\_scores\_communication | Categorical | Independent |
| review\_scores\_location | Categorical | Independent |
| review\_scores\_value | Categorical | Independent |
| reviews\_per\_month | Discrete | Independent |

One significant limitation of the dataset is the sparseness observed, with approximately 38% of entries missing for the 'price' variable and around 30% for 'reviews\_per\_month', 'review\_scores\_rating', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', and 'review\_scores\_location'. To mitigate this, a significant delimitation will be imposed, when entries with missing values for 'price', 'review\_scores\_rating', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', and 'review\_scores\_location' will be eliminated, while 'bedrooms' will undergo imputation. Null values in 'reviews\_per\_month' will be filled with zeros since they are likely indicative of listings with no reviews. Post-processing, the dataset will consist of 17,252 rows.

Additionally, the dataset's temporal scope is restricted to only one year, potentially limiting the ability to adequately capture long-term trends and seasonal variations. Moreover, the dataset may not encompass all relevant variables that influence Airbnb pricing, such as local events, economic indicators, or regulatory changes, thus constraining the comprehensiveness of the analysis.

Another important delimitation of the dataset is the selective inclusion of specific columns in the analysis, driven by assumptions about the factors likely to influence pricing in the Airbnb market. By focusing solely on the selected variables listed above, the analysis may overlook other potentially relevant factors that could impact pricing dynamics. This approach could limit the comprehensiveness of the model and its ability to capture the full spectrum of variables influencing Airbnb pricing in NYC.

**Data Gathering:** The data on Airbnb listings in New York City is available for free download in a compressed zip file containing a CSV file representing quarterly data for the last year. In its raw form, the dataset comprises 39,167 rows, capturing information on key variables such as pricing, location coordinates, room type, minimum nights, number of reviews, and review scores. The total data sparsity percentage is 15.4%, so some preprocessing steps--such as the steps outlined in the previous section--will be necessary to handle missing values. The data gathering process involves scraping and aggregating information from Airbnb listings, which are publicly available on the website. Additionally, columns that are deemed unnecessary for the analysis, such as host-related information, will be omitted to streamline the dataset. These steps are consistent with the best practices for Multiple Linear Regression analyses (Anderson et al., 2001).

**Data Analytics Tools and Techniques**: For this analysis, multiple linear regression will be employed to explore the relationship between Airbnb prices and various independent variables. The research question seeks to understand how factors such as location, room type, minimum nights, number of reviews, and review scores influence pricing, making multiple linear regression an appropriate technique. After preprocessing steps to manage null values, outliers, and encode categorical variables, Q-Q plots will be employed to assess the normality of the continuous variables' data distributions. Subsequently, an initial multiple linear regression model will be constructed to establish a baseline understanding of the relationship between the dependent and independent variables. To refine the model and identify the most influential variables, backward elimination will be applied, iteratively removing the least significant predictors. This iterative process ensures that the final model includes only the most relevant variables, enhancing its predictive power (Mishra et al., 2021). Once the final reduced model is established, residuals will be calculated to assess model fit, and metrics such as the residual standard error (RSE) and Mean Absolute Percentage Error (MAPE) will be computed to evaluate the model’s performance. RSE measures the average deviation of the observed values from the predicted values, providing an indication of how well the model fits the data, while MAPE calculates the percentage difference between the observed and predicted values, making it useful for interpreting the accuracy of predictions relative to the actual values (M, 2023). RSE focuses on the magnitude of errors, and MAPE provides insights into the relative accuracy of the model predictions. By considering both metrics, analysts can gain a comprehensive understanding of the model's performance and make informed decisions regarding its suitability for practical applications. The presentation layer will encompass a variety of visual tools, including univariate and bivariate graphs, Excel tables and charts, and a Tableau dashboard. These elements will serve to provide insights into the relationships between variables and aid in model interpretation. Additionally, recommendations for potential areas of improvement in Airbnb pricing strategies will be provided, based on the insights gleaned from the model's analysis.

**Justification of Tools/Techniques:** Python is a versatile programming language widely used for data analysis and machine learning tasks, making it a good choice for conducting multiple linear regression analysis. While R is also popular for statistical analysis, Python offers several advantages, particularly in terms of its superior processing time (ProjectPro, 2024). Considering the substantial dataset comprising 17,252 rows, careful consideration of processing time becomes paramount. Thus, Python was chosen over R for its efficiency in handling such large datasets.

Compared to SAS, Python offers a more flexible and cost-effective solution for conducting regression analysis. While SAS is renowned for its statistical analysis capabilities, it requires licensing fees and may have a steeper learning curve for users unfamiliar with its syntax (GeeksforGeeks, 2023). In contrast, Python is open-source and has extensive community support that makes it accessible to a broader audience. Moreover, Python's ecosystem of libraries such as scikit-learn, matplotlib, seaborn, and statsmodels provides comprehensive tools for regression analysis, visualization, and model evaluation.

The choice of libraries and packages for this analysis is deliberate and well-suited for the requirements of multiple linear regression. Pandas is utilized for data manipulation and preprocessing tasks, numpy for efficient numerical computation, and sklearn.preprocessing for feature scaling and encoding categorical variables. Matplotlib and seaborn are employed for data visualization, facilitating the exploration of relationships between variables and the discovery of patterns in the data. Statsmodels.api is used for building regression models and conducting statistical tests.

**Project Outcomes**: This analysis will produce a multiple linear regression model aimed at predicting the prices of Airbnbs in New York City. Upon successful model development, the key deliverable will be a comprehensive report detailing the methodology employed, the rationale behind variable selection, and the model's performance metrics. Support for the alternative hypothesis is found in a study called 'Online Stores: Analysis of User Experience with Multiple Linear Regression Model,' which determined that a multiple linear regression model effectively approximates the dependent variable by assessing the influence of predictor variables on sales data. (Acheme et al., 2021).

**Projected Project End Date**: May 3, 2024

**Sources**:

Acheme, D., Uddin, O., Makinde, A., Vincent, O., & Vincent, O. (2021). Online Stores: Analysis of user Experience with Multiple Linear Regression Model.

Anderson, D., Sweeney, D., & Williams, T. (2001). Contemporary Business Statistics with Microsoft Excel. Cincinnati, OH: South-Western. Retrieved on April 18, 2024, from <https://ruby.fgcu.edu/courses/tharring/80890/m3_1.htm>

Benítez-Aurioles, B. (2020). Impacts of the Peer-To-Peer Market on Tourist Accommodation on the Balearic Islands of Mallorca and Menorca. Island Studies Journal, 15(2), 353–370. <https://doi.org/10.24043/isj.108>

GeeksforGeeks. (2023, June 12). SAS vs R vs Python. Retrieved on April 15, 2024, from <https://www.geeksforgeeks.org/sas-vs-r-vs-python/>

Inside Airbnb. (n.d.). Get the data. Retrieved on April 11, 2024, from <https://insideairbnb.com/get-the-data>

M, Padhma. (2023, November 30). Evaluation Metric for Regression Models. Analytics Vidhya. Retrieved on April 12, 2024, from <https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/>

ProjectPro. (2024, April 18). Data Science Programming: Python vs R. Retrieved April 22, 2024, from <https://www.projectpro.io/article/data-science-programming-python-vs-r/128>

Petrovski, A., Petruseva, S., & Zileska Pancovska, V. (2015). Multiple linear regression model for predicting bidding price. Technics Technologies Education Management, 10, 386-393.

Mishra, C., Mohanty, L., Rath, S., Patnaik, R., & Pradhan, R. (2021). Application of Backward Elimination in Multiple Linear Regression Model for Prediction of Stock Index. In D. Mishra, R. Buyya, P. Mohapatra, & S. Patnaik (Eds.), Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies (Vol. 153). Springer, Singapore. <https://doi.org/10.1007/978-981-15-6202-0_56>

von Briel, D., & Dolnicar, S. (2021). The evolution of Airbnb’s competitive landscape. In S. Dolnicar (Ed.), Airbnb before, during and after COVID-19. University of Queensland. <https://doi.org/10.6084/m9.figshare.14195960>

**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.