METHODIST
COLLEGE | KUALA LUMPUR
Veritas Vincit Omnia

American Degree Transfer Program

Course: Programming with Python

Course Code: CSC 1620

Assignment Number & Title: COURSE PROJECT

Name of Lecturer: Yeoh Kuan Yew

Name(s):	**Student ID Number(s):**
Jeremy Lim Kin Lok	ADTP2208003

I/We declare that this is my/our own original work and any contributions made by others have been properly acknowledged and/or referenced.

Signed: *Jeremy*

Date of Submission: 05/04/2023

COURSE PROJECT

Program: American Degree Transfer Program

Semester: Jan 2023

Course Code & Course Title: Programming in Python with Lab

Project Marks: 100 (20% of Course Grade)

Airbnb Correlation Analysis

Analyze Airbnb dataset in performing correlation analysis using Python. This project aims to analyze the relationship variables for possible pricing and marketing strategy of unit rental.

The scope of analysis address two geographic coverages mainly at city area and at national level. Airbnb dataset is provided, scrapped from available units over 1 night stay on 30th April 2023.

Analysis requirements:

number of reviews, ratings and number of amenities.

2. Perform correlation analysis for possible relationships for unit rental ratings against number of beds, number of reviews, price and number of amenities.

Deliverable:

Produce a report detailing the results of the correlation analysis for price and ratings against other variables. Correlation plots (charts)

2. Proposal recommendation in ascertaining correlation relationship in establishing price for rental units on Airbnb.

Assumptions:

Data sample must satisfy a minimum sample size at 95% confidence level, with 5% margin of error. Use Sample Size Calculator (https://www.qualtrics.com/blog/calculating-sample-size/) in testing sampling size sufficiency for the respective variables.

2. Use Python libraries as required (Numpy, Seaborn, Pandas) along with other non-Python tools.

Project marks

a) Correlation analysis and Python scripting (70%)
b) Correlation report (30%)
c) Correlation plots (charts) in supplementing the correlation analysis (extra credit).

Submission Requirements:

a) Correlation analysis report and recommendation.
b) Python scripts in performing correlation analysis.

Due date: **4th May 2023, 2359**

Data Analysis Report on national & KL unit rentals

Introduction:

The goal for this Data Analysis Report is to examine the possible link between potential prices and their marketing elements for rental units.

The two geographic coverages that make up the analysis's scope are on a national and Kuala Lumpur centered coverage.

The data source (Yew, 2023) was scraped for April 30, 2023, from all accessible and available units for a one-night stay.

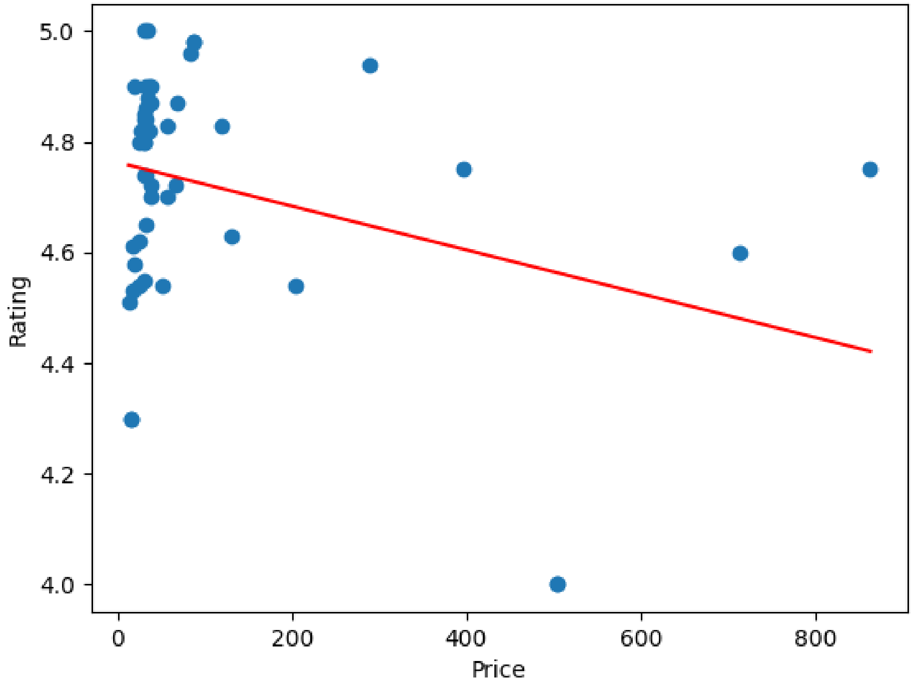
Data Analysis on Kuala Lumpur city area:

We would first display the Kuala Lumpur city area unit rentals (Total: 49).

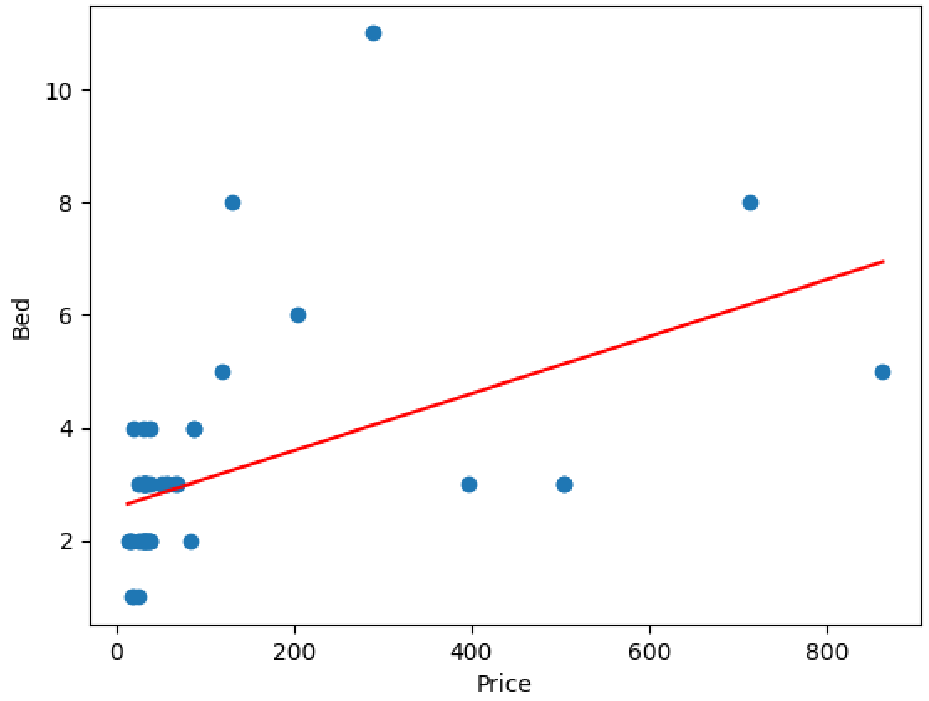
Format: [Graph generated]

[X against Y]:[Correlation value] ([Positive/Negative Linear Relationship])

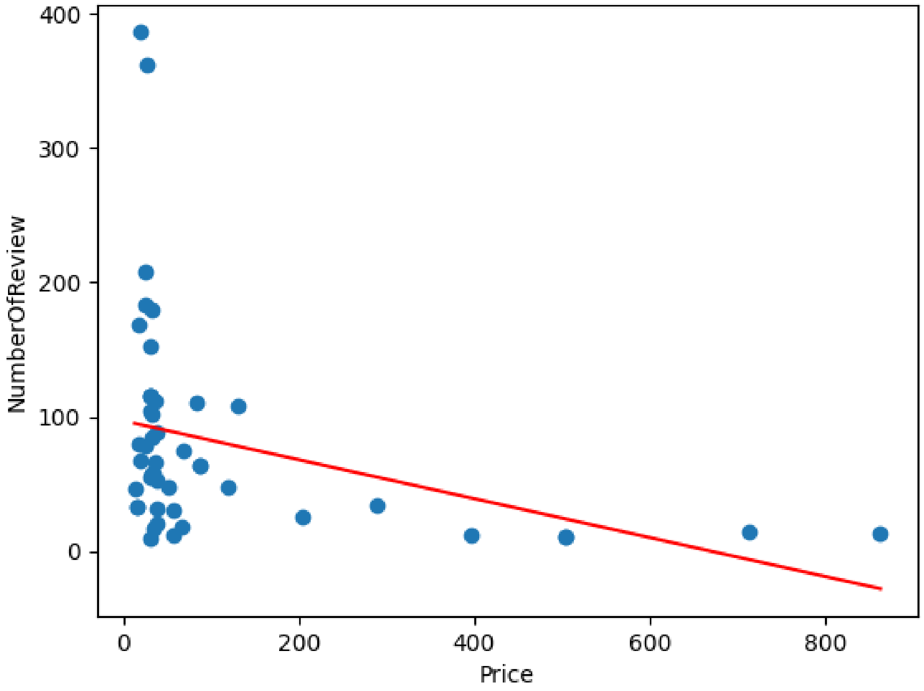
Comparing Prices with the other Variables:



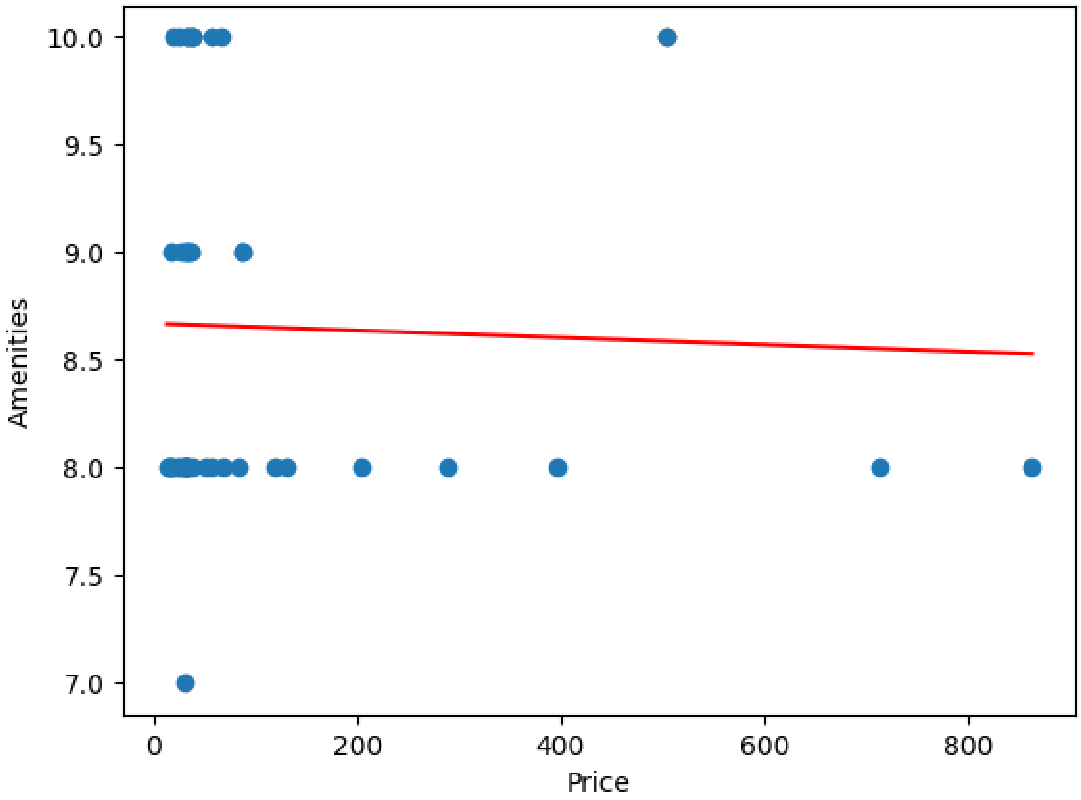
Price against Rating: -0.32187096234612667 (Negative Linear Relationship)



Price against Bed: 0.4959964390446656 (Positive Linear Relationship)

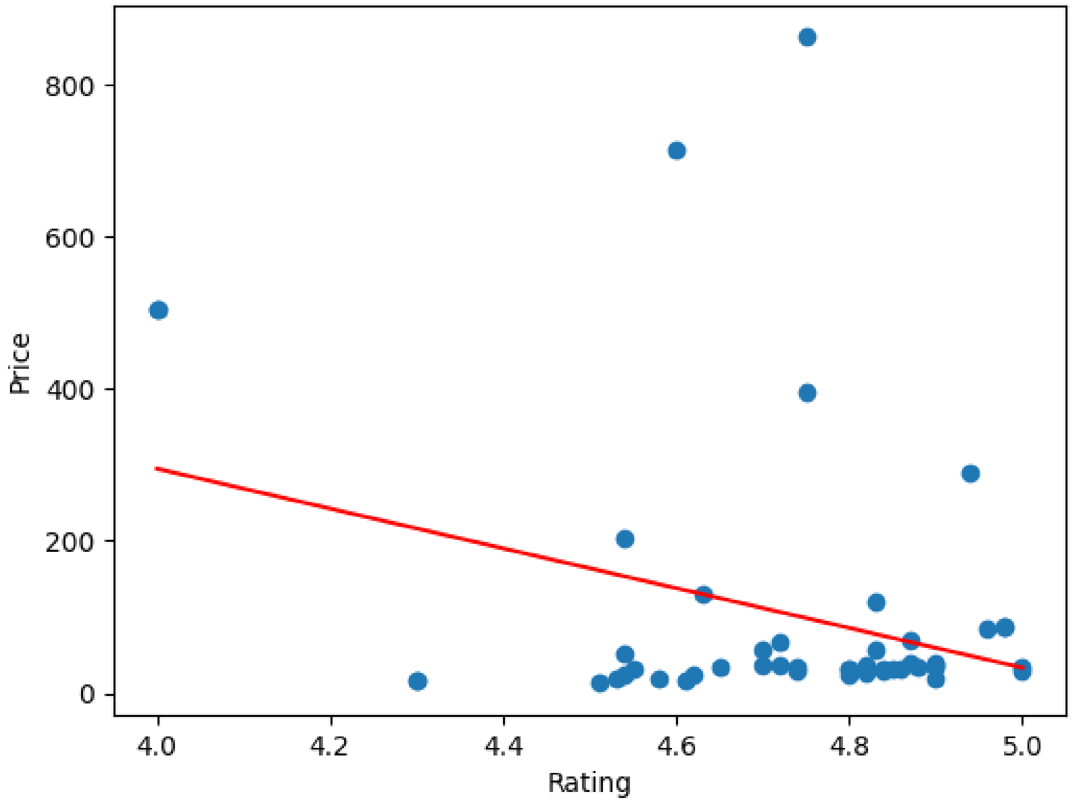


Price against NumberOfReview: -0.33536440584857974 (Negative Linear Relationship)

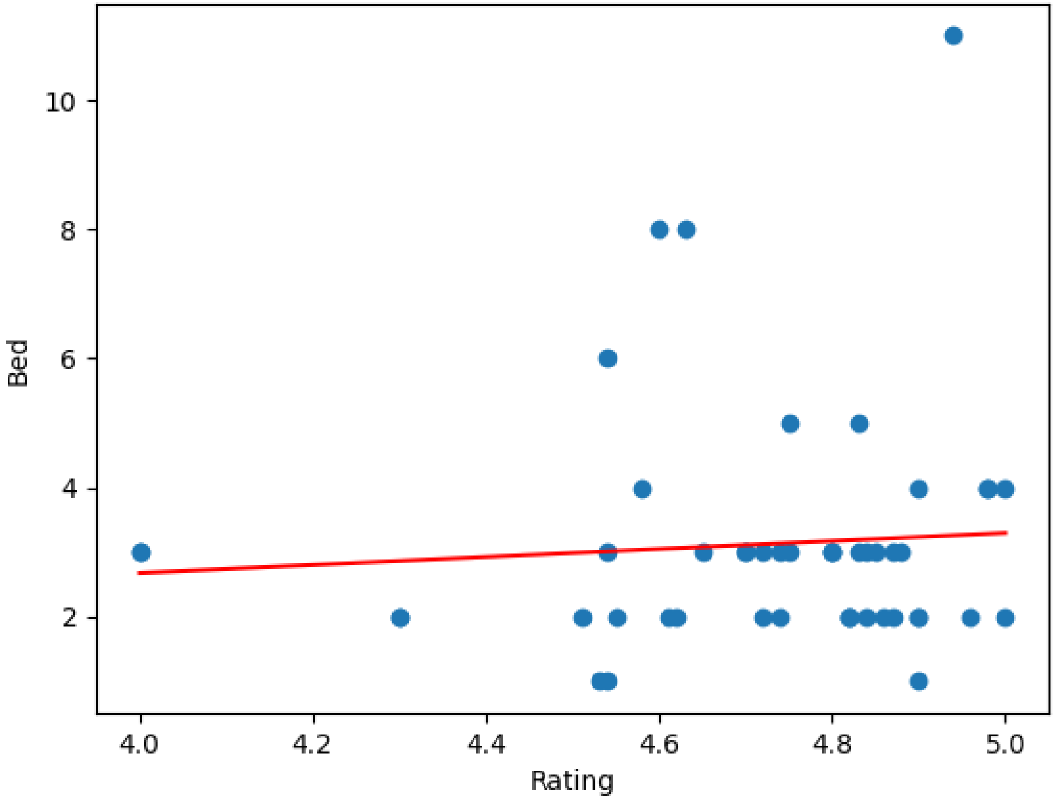


Price against Amenities: -0.034015842344987404 (Negative Linear Relationship)

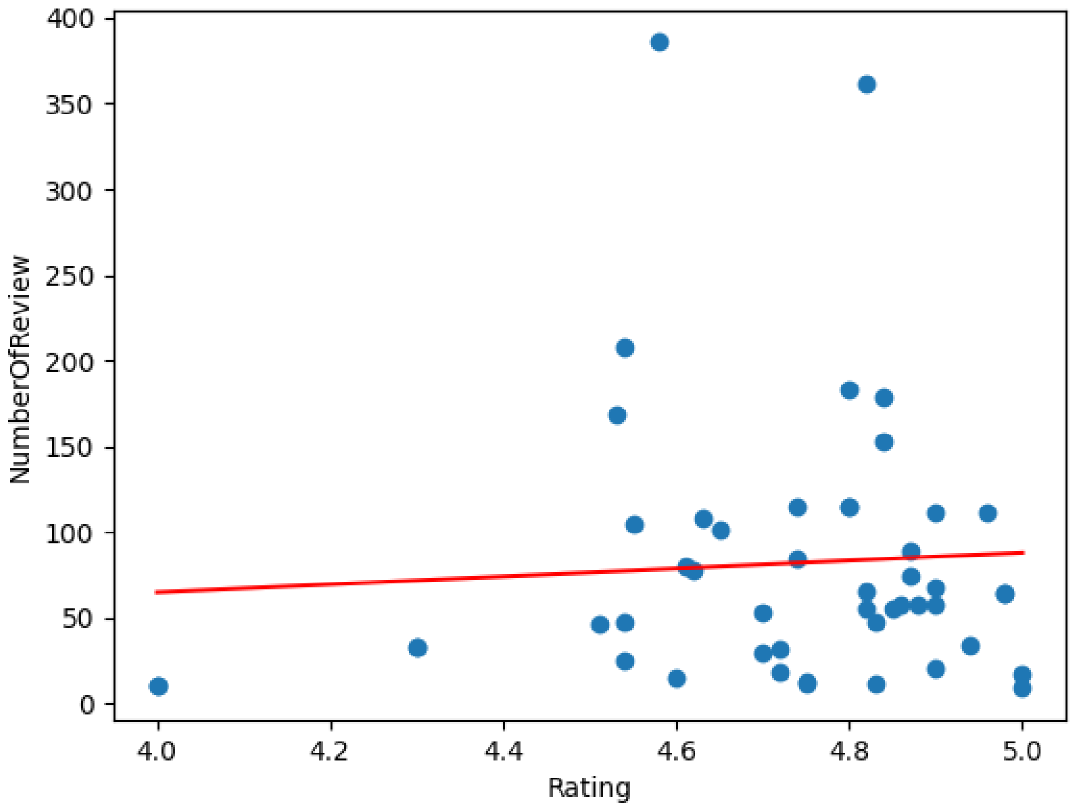
Comparing Rating with other Variables:



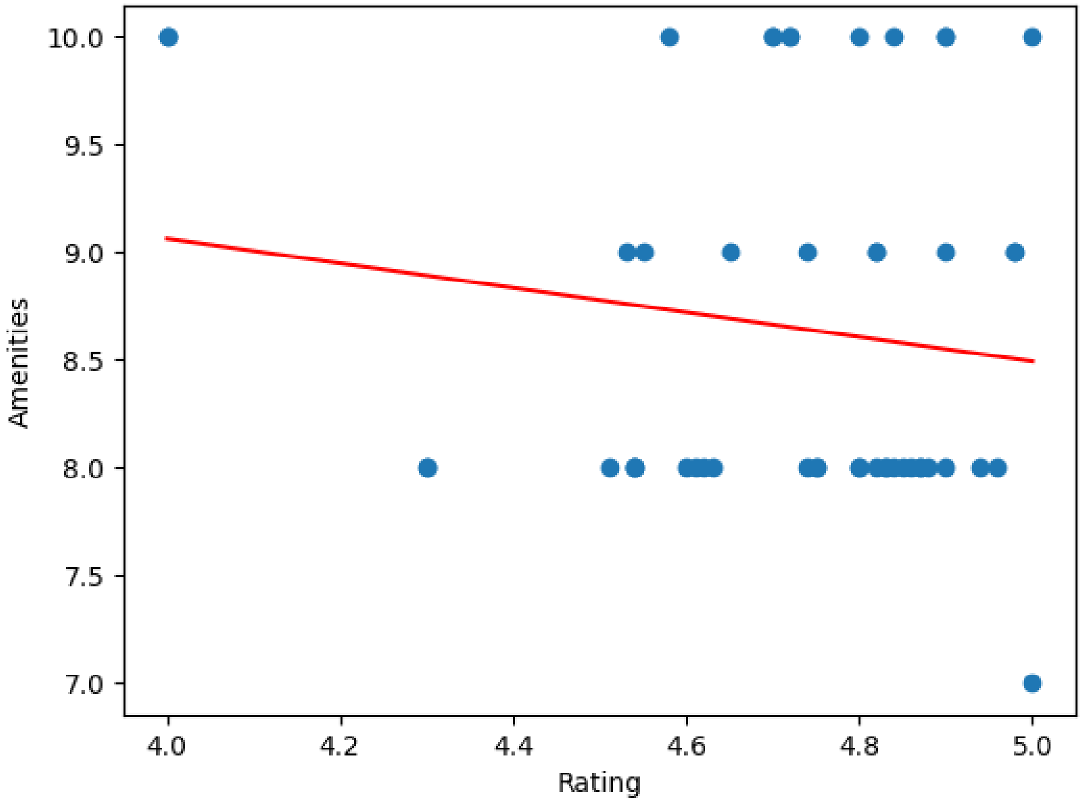
Rating against Price: -0.32187096234612667 (Negative Linear Relationship)



Rating against Bed: 0.07454875866425979 (Positive Linear Relationship)



Rating against NumberOfReview: 0.06597907303810822 (Positive Linear Relationship)



Rating against Amenities: -0.1449803994229948 (Negative Linear Relationship)

Data Analysis on a National Level:

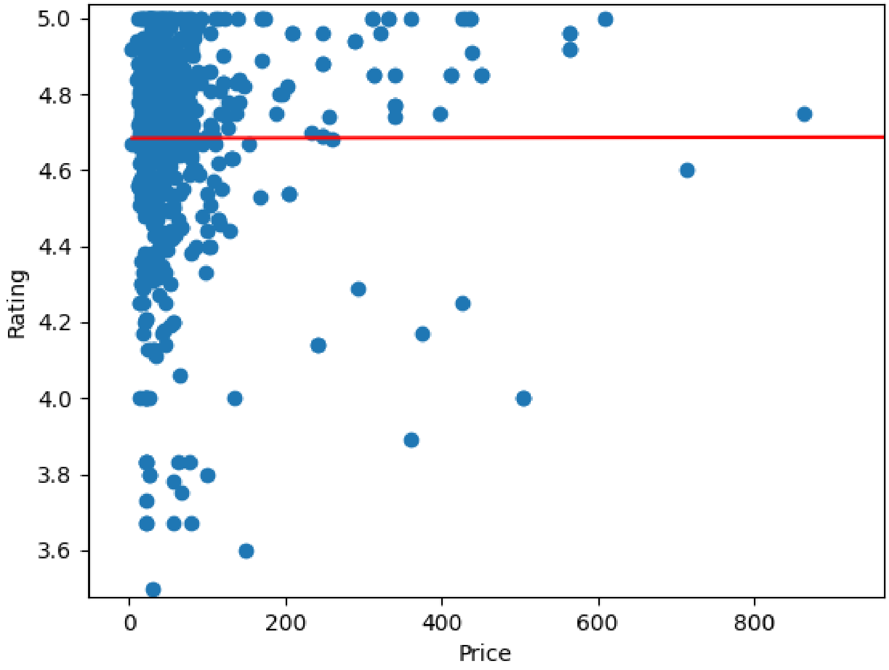
We would now display the National Coverage unit rentals (Total: 625).

When comparing Prices with the other variables, we will focus onto the majority cluster due to having an absurd price of approximately $21000+.

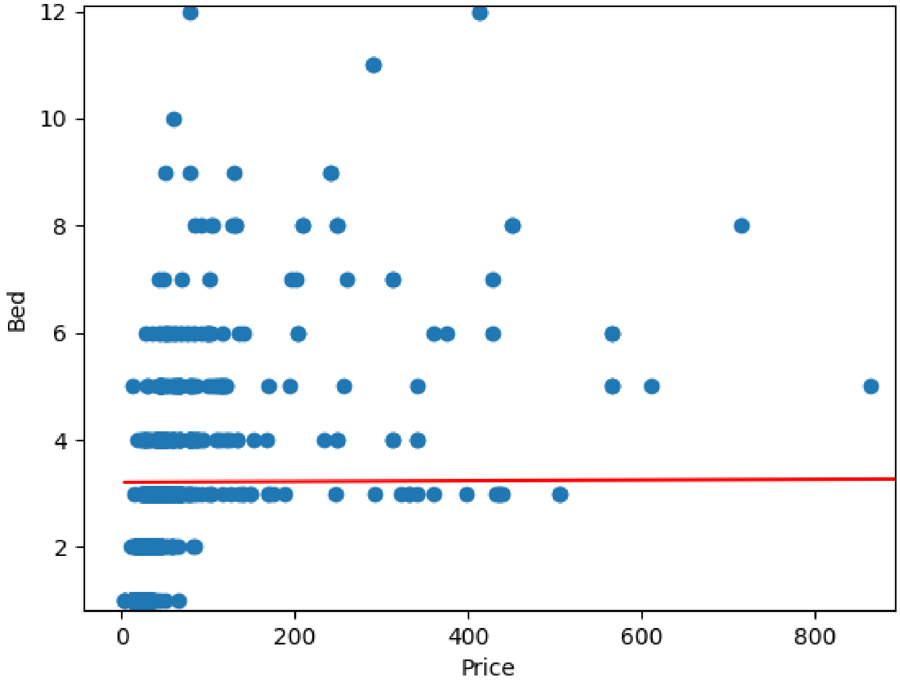
Format: [Graph generated]

[X against Y]:[Correlation value]([Positive/Negative Linear Relationship])

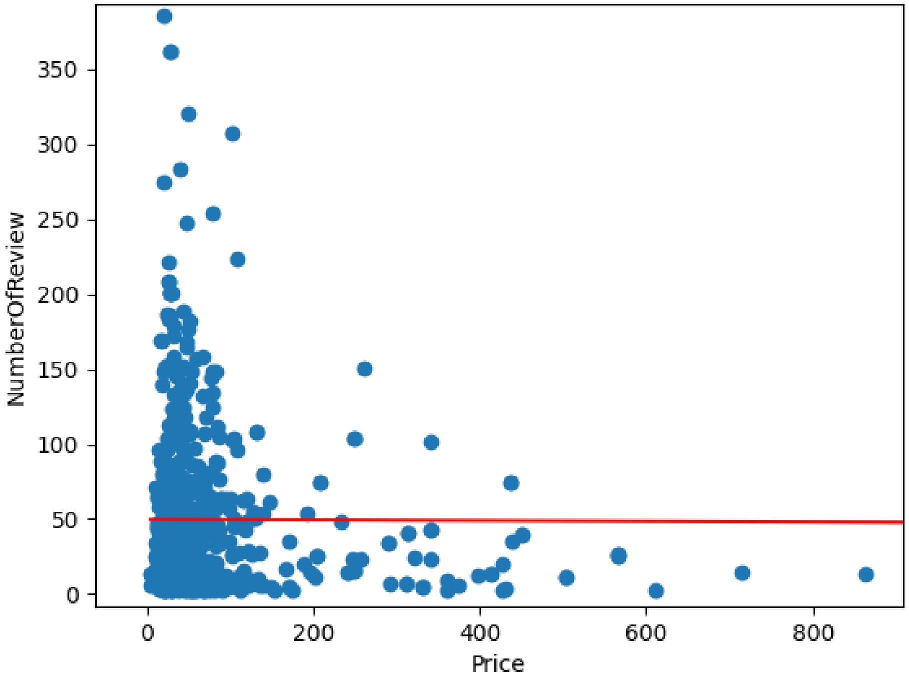
Comparing Prices with the other Variables:



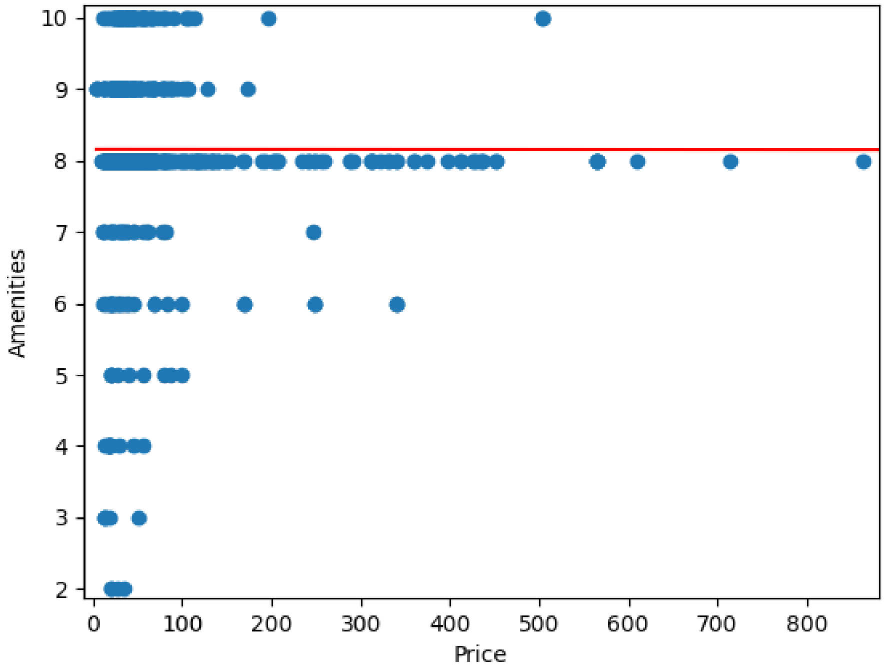
Price against Rating: 0.011556037859860882 (Positive Linear Relationship)



Price against Bed: 0.039776104604366314 (Positive Linear Relationship)

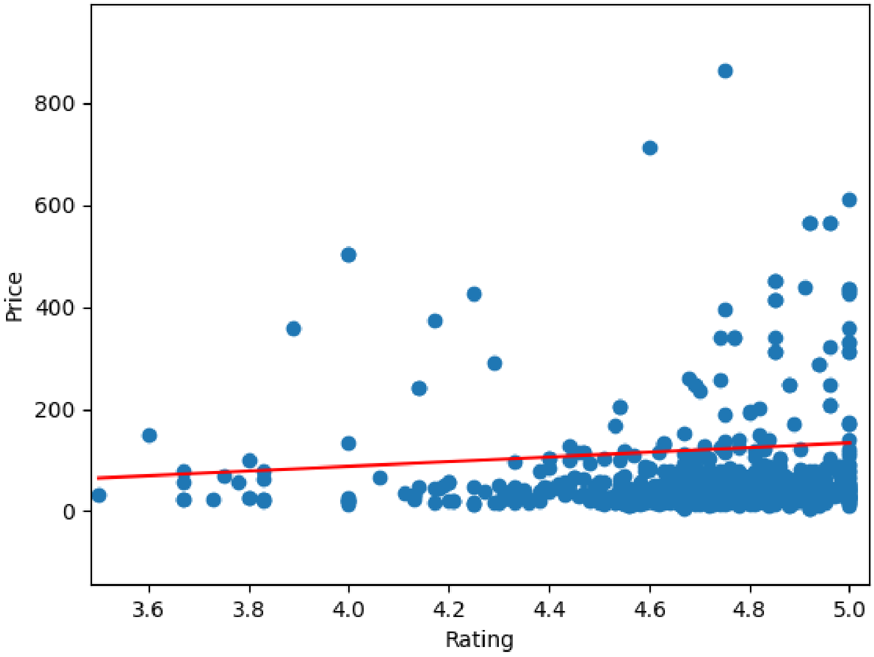


Price against NumberOfReview: -0.04146420835938946 (Negative Linear Relationship)

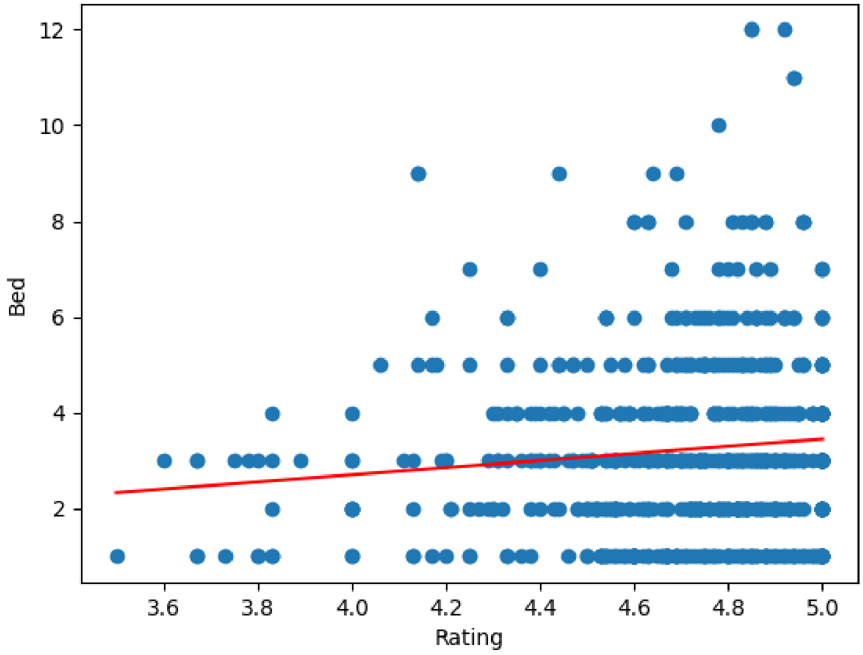


Price against Amenities: -0.005658620520513276 (Negative Linear Relationship)

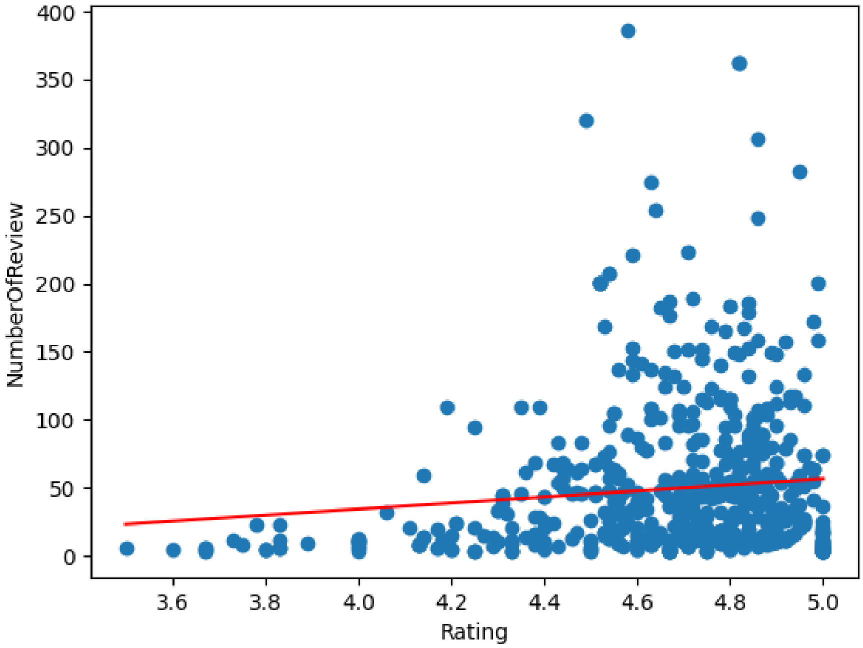
Comparing Ratings with the other Variables:



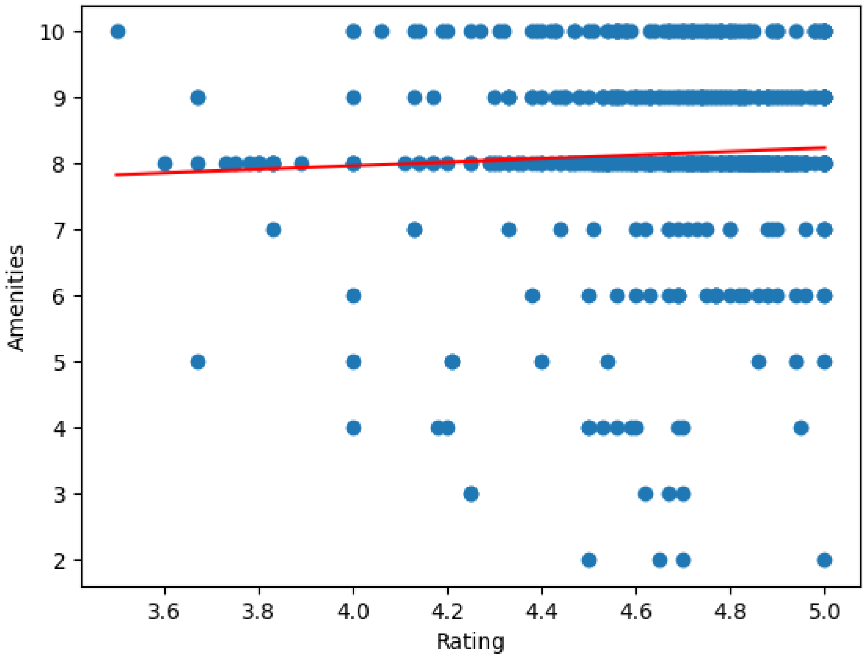
Rating against Price: 0.011556037859860882 (Positive Linear Relationship)



Rating against Bed: 0.10690849502743276 (Positive Linear Relationship)



Rating against NumberOfReview: 0.11026445829740737 (Positive Linear Relationship)



Rating against Amenities: 0.055819643389421865 (Positive Linear Relationship)

Findings for both Kuala Lumpur and National coverage:

Kuala Lumpur city area:

When it comes to comparing prices with other variables in Kuala Lumpur assets, majority of the graphs are negative linear relationships except for bed.

However, since the average American family size of today is 3.13 (Duffin,2022), it is safe to assume most clients need less than 5 beds per booking. Thus, we can partially ignore the bed's positive linear relationship.

This means for a client with the best saving money as his/her best interest in mind, it is still viable for him/her to pick the cheapest room without getting setbacks.

When it comes to comparing ratings with other variables in Kuala Lumpur assets, it can be said that those with more beds and a greater number of reviews obtain higher ratings. Also, these rental rooms with high ratings will cost less. The only drawback is that they have less amenities.

It can be viewed that rooms receive a better rating by increasing the number of beds, lowering the price and having a greater number of reviews obtain. It seems Amenities is not preferable maybe because tourist who rent the rooms prefer to go outside and see the landscape/area inside of staying in the indoor rental units.

National level:

When it comes to comparing prices with other variables in National level assets, it seems that there is a positive increase in ratings and beds as the price increases. The only problem is that the number of reviews and amenities decreases as the price increases.

A possible reasoning behind the number of reviews is that less people can afford to rent expensive rooms and determine whether the luxury was worthwhile. The downfall in amenities and increase in ratings might be because the amenities are high-tech and cool but more expensive.

When it comes to comparing ratings with other variables in National level assets, surprisingly all the graphs comparing the values have a positive linear relationship. This means that the highest rated accommodation rooms have more expensive prices but better number of reviews, beds, and amenities.

This could mean that if a client wants the safest choice when it comes to renting an Airbnb nationwide, he/she should pick the ones with the best reviewed unit as it grantees the best accommodations.

Conclusion / Recommendation:

Despite being able to abstract potential and meaningful data from the findings, it is recommended that no further action should be taken place as of yet. This is because of the poor number of samples we have at our disposal. Despite originally having 2423 National data and 152 Kuala Lumpur data at my disposal, I was only able to use 625 National data and 49 Kuala Lumpur data due to the missing gaps of knowledge and 0s inside the data file. As a result, only 29% of the data can only be use. This result in a lack of data size and a lack of adequate sample size of around 100s to 1000s (Versta Research, n.d.).

Due to this I believe more data is needed to be scavenged from multiple different Airbnb websites that contain full reliable information unlike the Airbnb

dataset is provided to me.

Last but not least, when trying to find the most optimal unit rental there is out there, it is better to have more variables such as 'Tourist location spot', 'Number of transports', 'Number of Malls', 'Number of Food stalls/shops', and 'Number of police stations'.

Reference:

Duffin, E. (2023, December 12). *Average number of people per family in the United States from 1960 to 2022*. Statista. https://www.statista.com/statistics/183657/average-size-of-a-family-in-the-us/

Versta Research. (n.d.). *Statistically Significant Sample Sizes*. https://verstaresearch.com/blog/statistically-significant-sample-sizes/

Yew, Y., K. (2023, April 28). *MY_details.v2.csv*. Excel.