NY Rental Properties Pricing:
Feature Reduction and Visualization

Brandon Palomino, Yosha Mundhra
MATH 250, Dr. Chen
05/19/2023

TABLE OF CONTENTS

I DATA BACKGROUND AND TRANSFORMATION

New York City (NYC) is a highly sought-after rental market that attracts residents who are planning to seek out a living domain. These rental prices can vary significantly based on many different components such as area, amenities, and proximity to other attractions. With many variations in rental areas such as apartments to hotels, there are differences in prices throughout the NYC area. With rental prices fluctuating commonly due to economic trends and real estate development, securing a good deal in this competitive market is crucial. Exploring rental property pricing data from NYC should provide a better understanding on why these patterns occur.

The data set used comes from Kaggle and is called NY Rental Properties Pricing. The data set consists of 17,614 observations of rental properties. The dataset has a total of twelve features, however only ten of the twelve were relevant since key value and id hold no significance to differentiating the data. Of the ten features, two are categorical while the rest are numerical. The data set contains rental property observations in 2022 from the five boroughs of New York City. The data set includes information on latitude, longitude, price, availability, reviews, and more.

Throughout the analysis on the rental properties pricing dataset, MATLAB was used as the main tool to process the data set and visualize them under several dimensionality reduction techniques. The toolbox Stats is used throughout each model as well.

In regards to feature scaling, some, but not all features differ by an order of magnitude. Due to this, transformations were applied to preserve the geometry and reduce the effects of distortion across all the data. To ensure uniformity in data for comparison, normalization was applied. In order to reduce skewness in data, log transformations were applied.

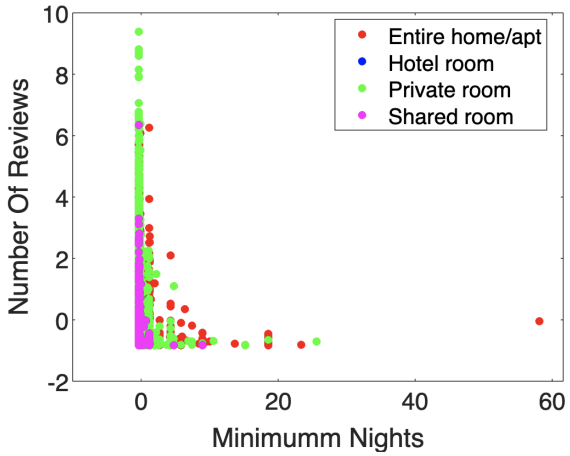

Figure 1: Before Log transformation

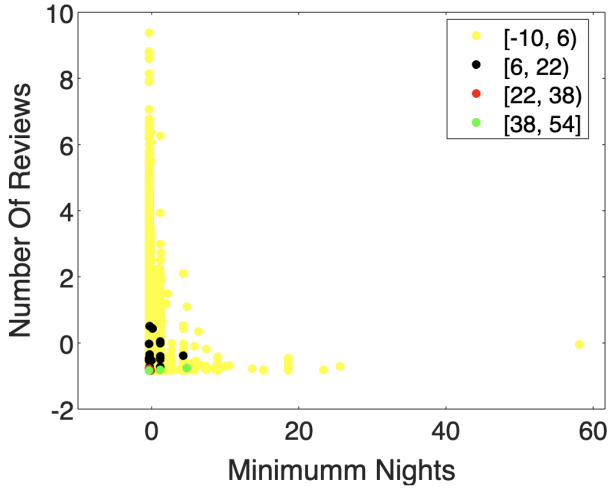

Figure 2: Before Log transformation

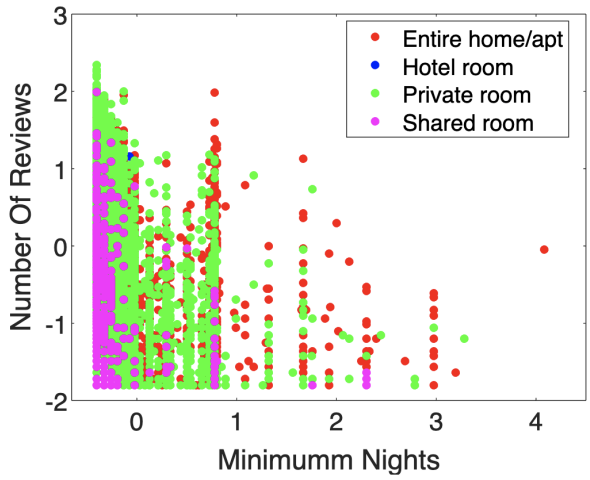

Figure 3: After Log transformation

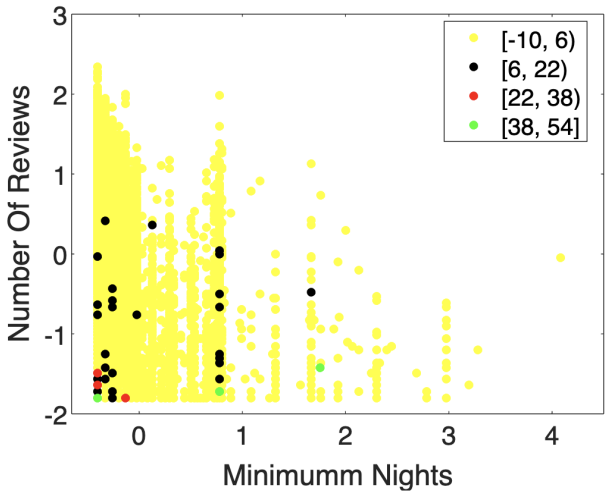

Figure 4: After Log transformation

II DATA VISUALIZATION

Data must first be visualized and its features understood before dimensionality reduction can be applied. We initially chose to plot the data categorically by room type because there were so many features. Later, we categorically plotted the data according to a price range. A rental home or apartment may be categorized as an entire home or apartment, hotel room, private room, or shared room. The four categories of rooms are used in figure (5) to group the data. The x-axis represents the different

categories of room types in the dataset, and the y-axis represents the frequency or number of occurrences of each room type. The plot also shows that the distribution of room types is skewed towards "Entire home/apt" and "Private room" categories, which are the most common room types in the dataset. This indicates that the dataset contains more listings for entire homes/apartments and private rooms compared to shared rooms.

Figure (6) shows a word cloud based on room types where the size of each word in the word cloud reflect the frequency or significance of that particular room type and its associated price.

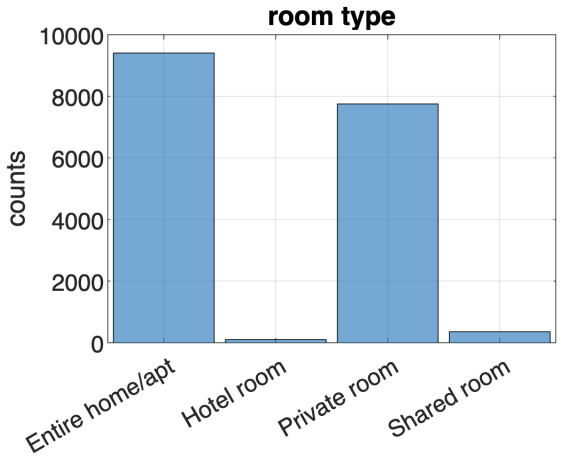

Figure 5: Rental property by room type

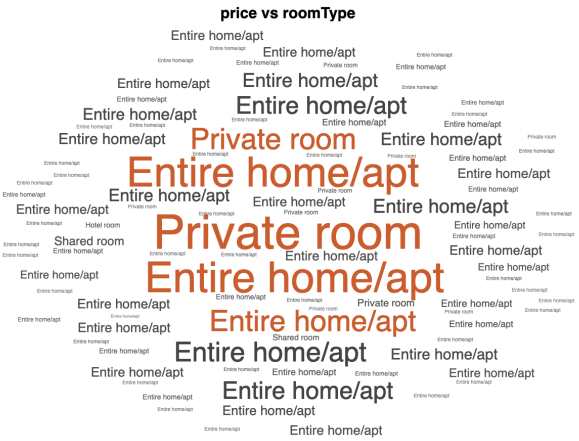

Figure 6: Word cloud of room types

For more context on the data, the data was also grouped by neighborhood as well. Figure 3 shows there are different rentals based on this category. However, since there are many subcategories that each makeup less than one percent of the data, dimensionality reduction was not applied in regards to neighborhood. The categories of neighbourhood are used in figure (7) to group the data in the form of a pie chart. Figure (8) shows a word cloud based on neighbourhood where the size of each word in the word cloud reflect the frequency or significance of that particular neighbourhood and its associated price.

Looking at Figure (9), each rental property was plotted on a graph in relation to longitude by latitude. Across all points, there does not seem to be a form of separation among all points, however, private

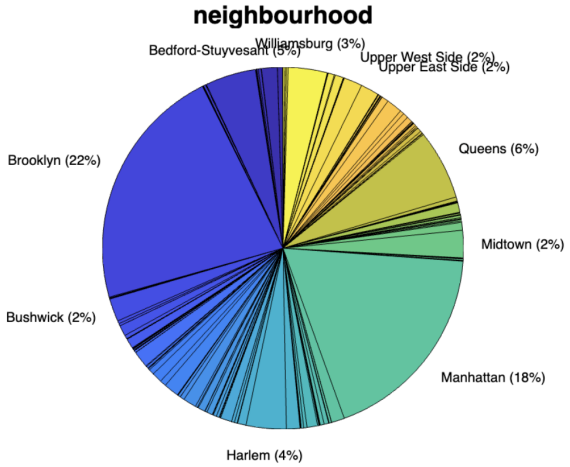

Figure 7: Rental property by neighborhood

Figure 8: Word cloud of neighborhoods

rooms and entire homes/apartments are scattered evenly. Shared rooms however make up the manhattan area of the plot, which makes sense given how dense the area is at that location. Figure (10) does the same but with categorization by price instead. Based on the pricing plot, a majority of rental properites have the same lower relative price range. However, some properties in the Manhattan have higher prices than the norm.

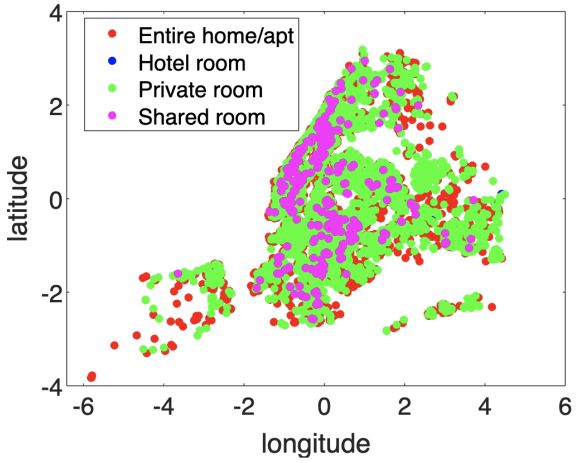

Figure 9: Longi. by Lati. by Room Type

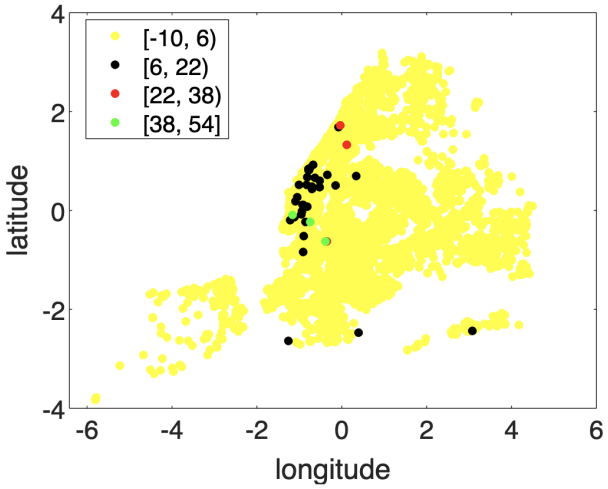

Figure 10: Longi. by Lati. by Price Range

Figure (11),(12),(13), show the variation in Number of reviews, Reviews per month and Price based on Minimum number of nights grouped

by room type. Most of the overlap is seen for "Entire home/apt" and "Private room" categories. Shared room has lower minimum number of nights.

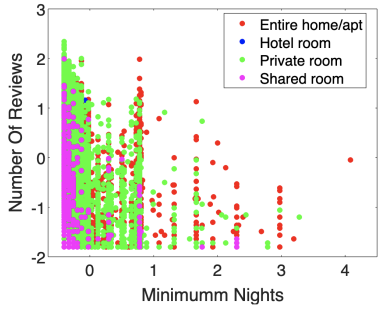

Figure 11: No. of reviews Vs Minimum nights

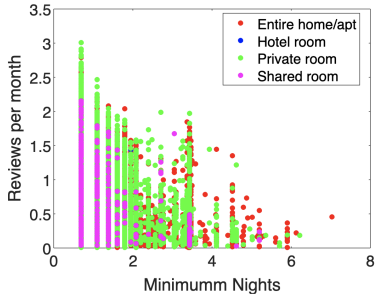

Figure 12: Reviews per month Vs Minimum nights

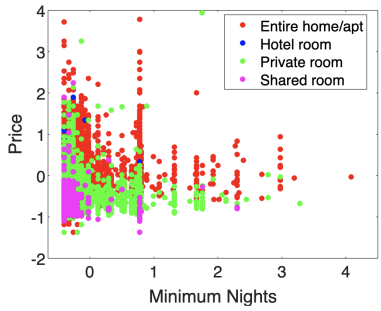

Figure 13: Price Vs Minimum nights

Figure (11),(12),(13), show the variation in Number of reviews, Reviews per month and Price based on Minimum number of nights grouped by price range. It demonstrates that most of the rental properties fall in the lower price range irrespective of the room types and there are fewer outliers with the higher price range.

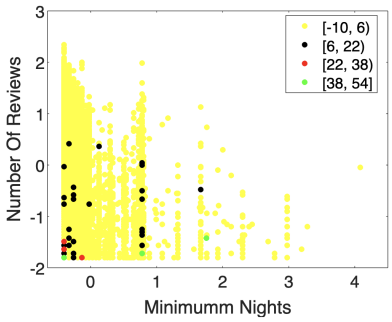

Figure 14: No. of reviews Vs Minimum nights

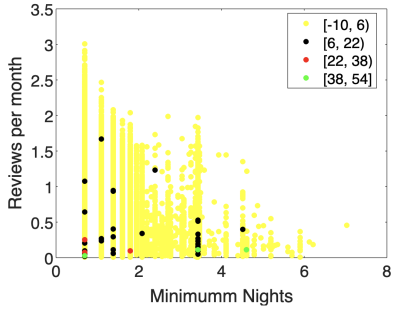

Figure 15: Reviews per month Vs Minimum nights

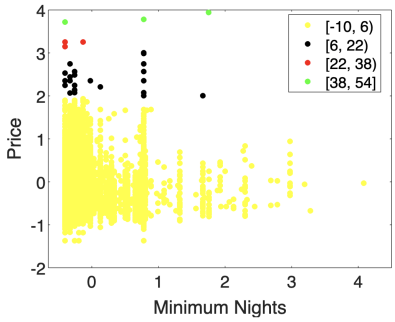

Figure 16: Price Vs Minimum nights

Figure (17),(18),(19), show the variation in Number of reviews, Reviews per month and Price with each other grouped by room type.

Figure (20),(21), show the variation in Number of reviews, Reviews per month and Price with each other grouped by price range.

Figure (22), show the variation in Price for the four different room types grouped by price range. "Entire home/apt" and "Private room" are among the ones with the higher price range.

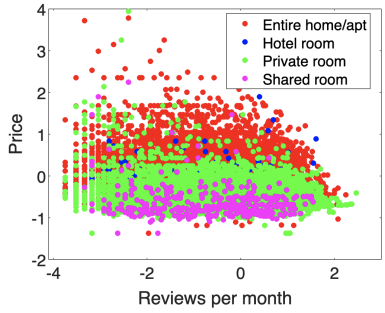

Figure 17: Price Vs Reviews per month

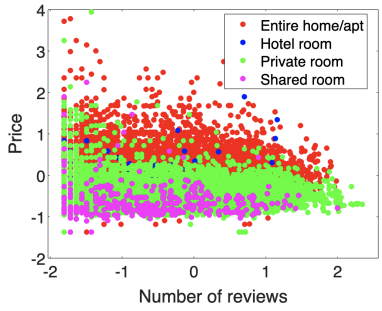

Figure 18: Price Vs No. of reviews

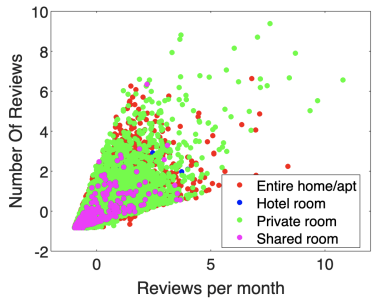

Figure 19: No. of reviews Vs Reviews per month

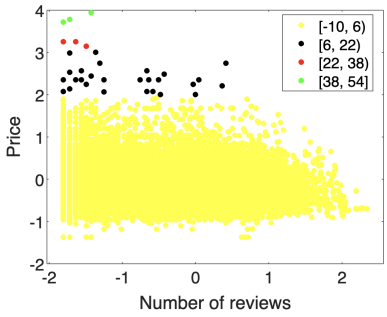

Figure 20: Price Vs No. of reviews

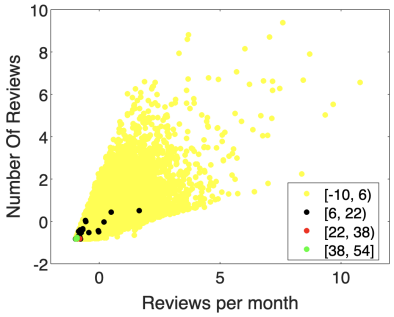

Figure 21: No. of reviews Vs Reviews per month

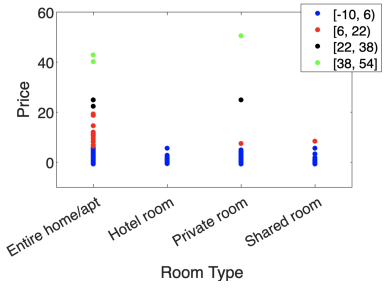

Figure 22: Price Vs Room types

Figure (23),(24), (25) show the variation in Price based on the availability in 2020 and days occupied in 2019 grouped by room type. We see that "Entire home/apt" and "Private room" were available and occupied most of the time shared room and hotel room were less available and occupied. Less occurrence of these two can also be one of the reasons for this visual.

Figure (26),(27), (28) show the variation in Price based on the availability in 2020 and days occupied in 2019 grouped by price range. The lower price range properties were available and occupied most of the time compared to the higher range properties.

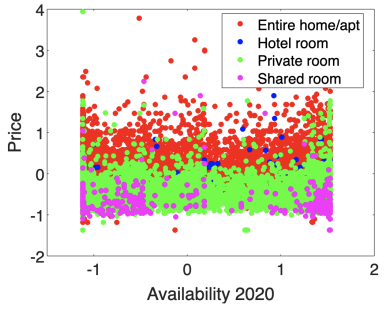

Figure 23: Longi. by Lati. by Price Range

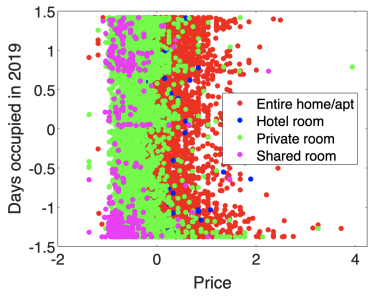

Figure 24: Longi. by Lati. by Room Type

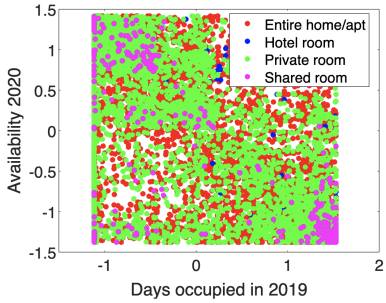

Figure 25: Longi. by Lati. by Price Range

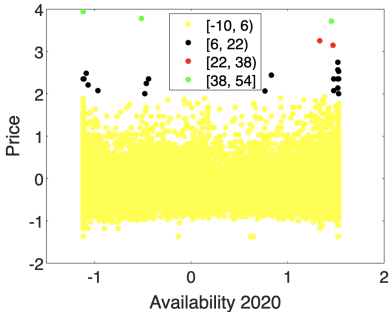

Figure 26: Longi. by Lati. by Room Type

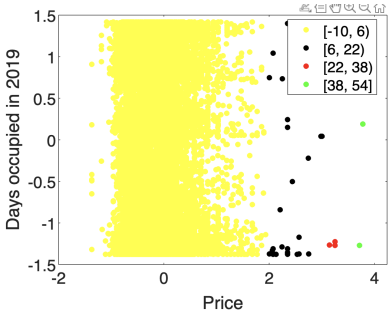

Figure 27: Longi. by Lati. by Price Range

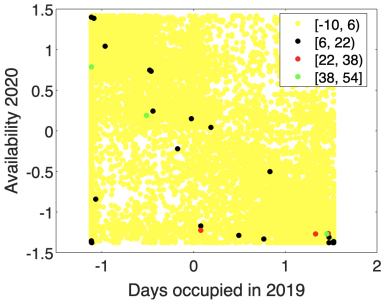

Figure 28: Longi. by Lati. by Price Range

III PCA

The goal of PCA is to compute new variables by applying linear transformation to the original observed data and project the new variables along the direction of maximum variability explained by them onto a 2D coordinate system, where the axes stand in for the directions of the data with the highest levels of variation. Examining the projected data helps to distinguish distinct groupings that are formed in the 2D projection space orthogonal to each other. Figures (29) and (30) show the 2 dimensional view of the data based on the Room types and Price range respectively after applying PCA. We see a lot of overlap for the category Room type. It might be because of the similar price range, days occupied or similar geographic location. The second plot based on the price range demonstrates that most of the rental properties fall in the lower price range irrespective of the room types and there are fewer outliers with the higher price range.

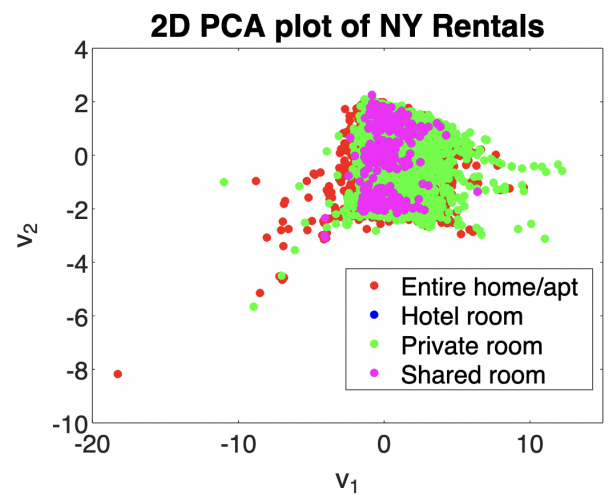

Figure 29: PCA - Room types

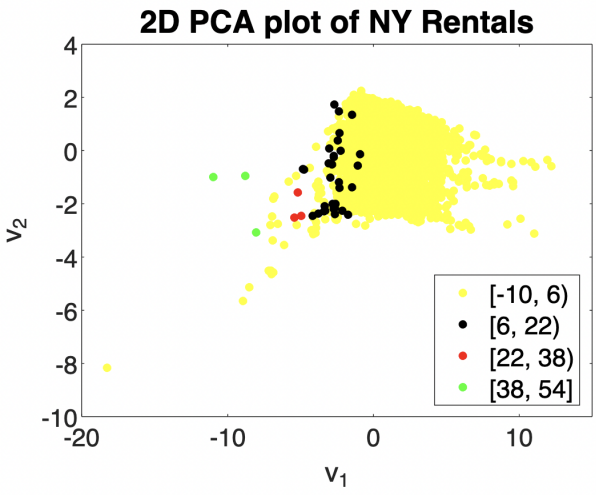

Figure 30: PCA - Price range

Figure (31), (32), (33) and (34) show the pairwise plots for "Private room , Shared room", "Entire home/apt , Hotel room", "Entire home/apt , Shared room", and "Hotel room , private room" respectively. We notice significant overlap in all four of the plots.

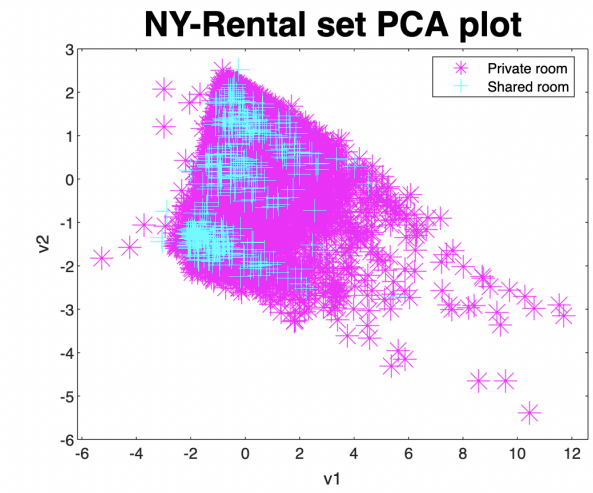

Figure 31: PCA pairwise - Private room , Shared room

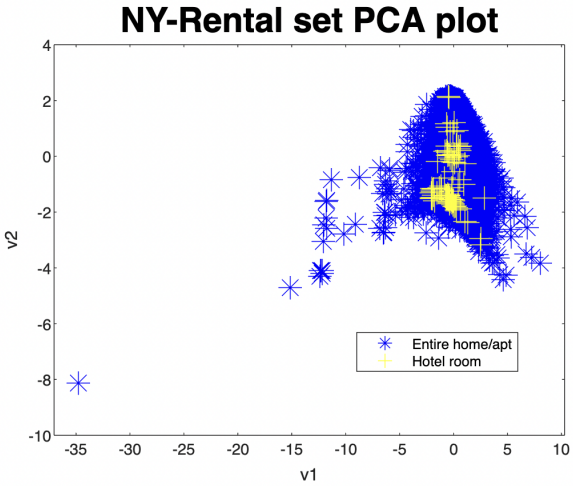

Figure 32: PCA pairwise - Entire home/apt , Hotel room

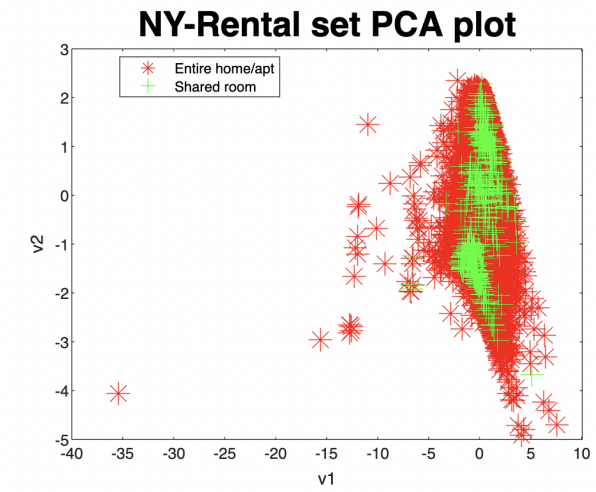

Figure 33: PCA pairwise - Entire home/apt , Shared room

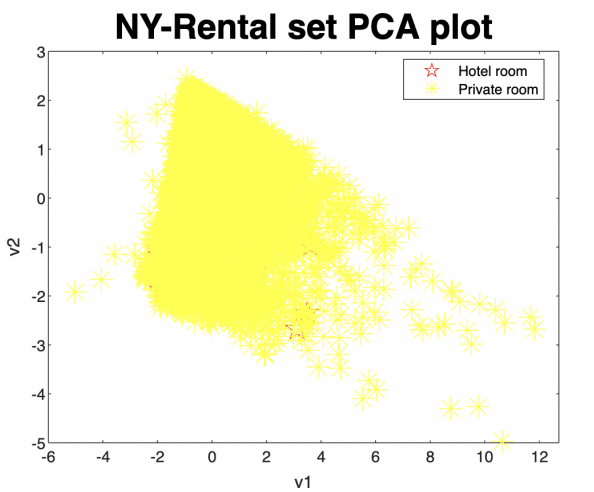

Figure 34: PCA pairwise - Hotel room , private room

IV LDA

The goal of LDA is to compute new variables by applying linear transformation to the original observed data and project the new variables onto a 2D coordinate system to maximize the separation between the groups, where the axes stand in for the directions of the data with the highest levels of variation. Figures (35) and (36) show the 2 dimensional view of the data based on the Room types and Price range respectively after applying LDA. PCA was applied to the data before LDA. We again see a lot of overlap for the category Room type. LDA does not perform significantly better than PCA.Similar price range, days occupied or similar geographic location can be one of the reason to explain this result. The second plot based on the price range demonstrates that most of the rental properties fall in the lower price range irrespective of the room types and there are fewer outliers with the higher price range. We can see clear distinction in 3 of the categories but there is still overlap between the two lower price ranges.

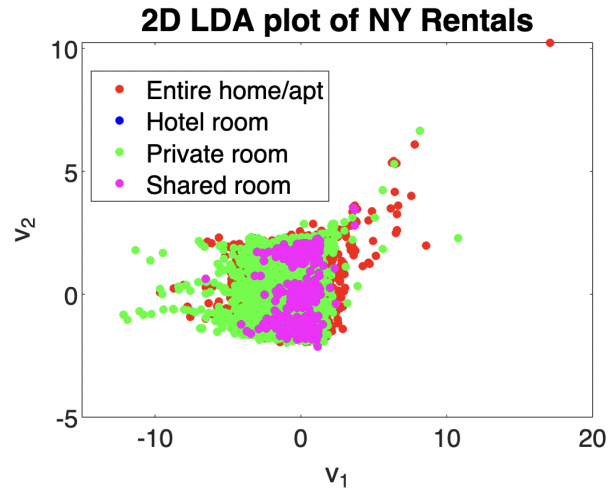

Figure 35: LDA - Room type

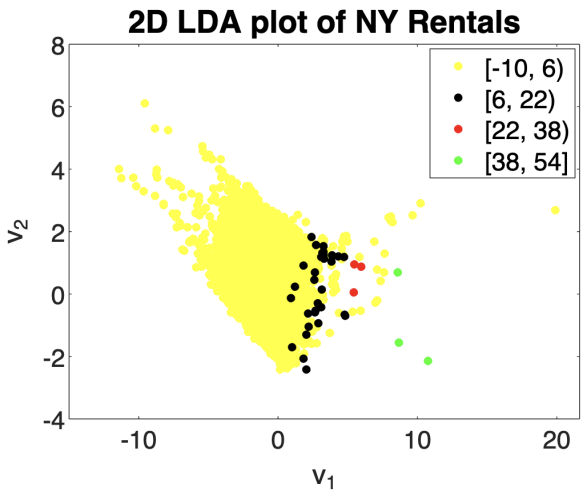

Figure 36: LDA - Price range

V MDS

The goal of multidimensional scaling (MDS) is to project high dimensional data down to only two dimensions while preserving relative distances between observations. For the purposes of the rental pricing data, cosine metric was used initially as a means to create pairwise similarities between the data. However, after examining the data at a closer glace, cityblock metric was used instead since it provides a more interesting take on the rental data. The results of applying MDS with a cosine dissimilarity to the data are poor, as indicated by a Kruskal Stress of about 0.21 for both room type and pricing reductions in Figures (37) and (41). Looking at the plots for cosine metric, we can see that there is extreme overlap among all the data points regardless of what room type or pricing range they were in. That is, approximating and projecting angles from a high-dimensional space into these metric spaces is not a good fit for the data - this suggests, perhaps, that the structure of the data is simpler, meaning that a different metric should be used instead. Thus, when applying the city block metric for MDS, there seems to be a star shape plot when looking at the 3D and 2D plots for MDS as shown in Figures (39) and (43). Looking closer at the new plots, the MDS projections using the cityblock metric are interesting in that they seem to agree with the PCA projec-

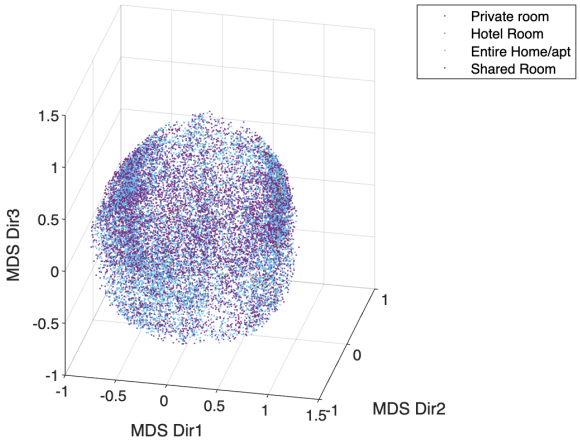

Figure 37: 3D MDS Cosine Metric

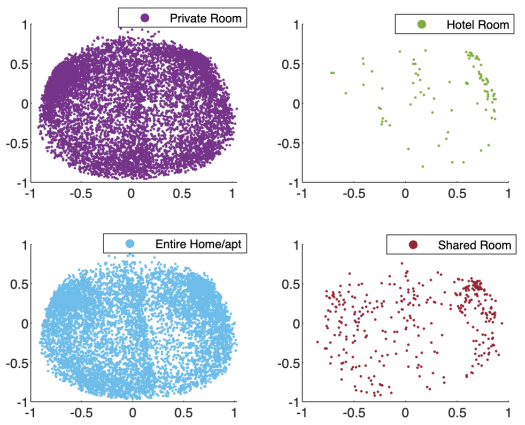

Figure 38: 2D MDS Cosine by Room Type

tions shown earlier. While classes are not able to be distinguished easily, groups of classes possessing similar numerical features are. This seems to imply that when general distances from the original space are kept, we get poor separation across classes but good separation between groups with identical numerical properties. When we start to take into account in-class and out-of-class differences, it appears that we obtain good class separation with no significant differences between general groups of data.

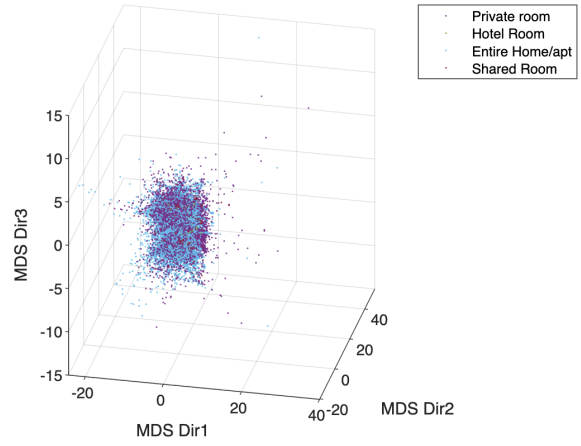

Figure 39: 3D MDS Cityblock Metric

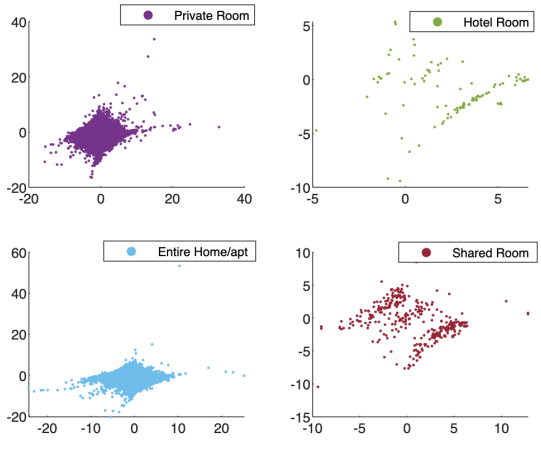

Figure 40: 2D MDS Cityblock by Room Type

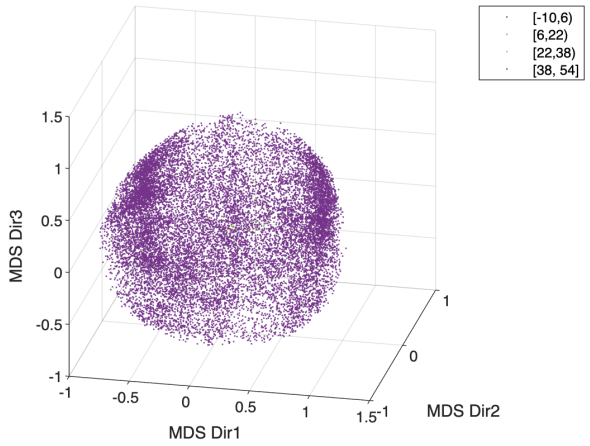

Figure 41: 3D MDS Cosine Metric

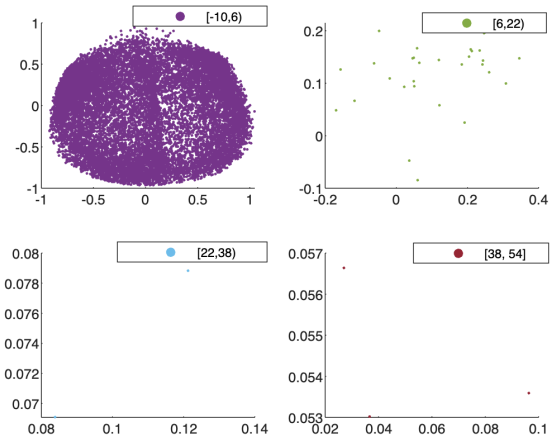

Figure 42: 2D MDS Cosine by Price

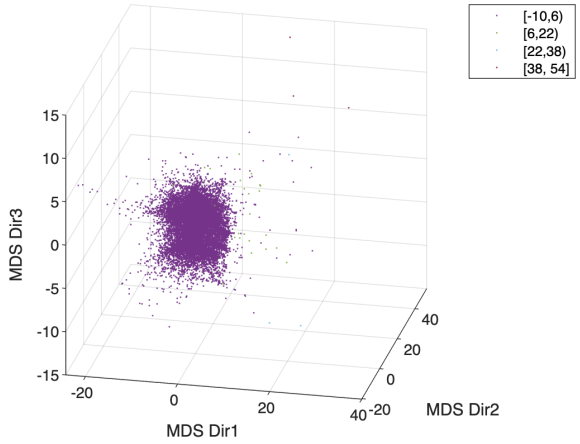

Figure 43: 3D MDS Cityblock Metric

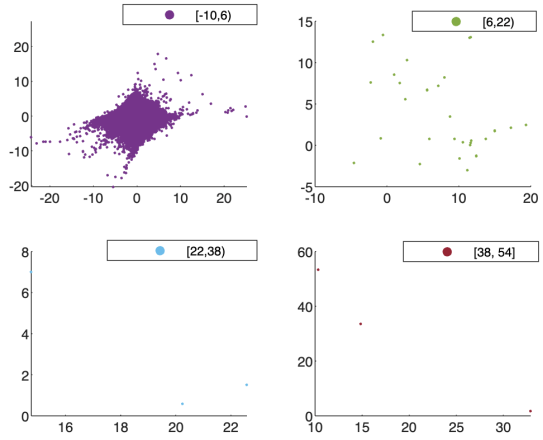

Figure 44: 2D MDS Cityblock by Price

VI Conclusion

In conclusion, all dimension reduction strategies demonstrated some separation between pricing and room categories. Throughout each dimensionality reduction technique, most data points overlapped with each other in specific areas on the plot. When comparing the findings, we can see that each room type and price is broken down according to a similar shape or curve. In conclusion, it may be said that all rental properties have features in common. They can all be categorized differently, though, based

on other variables.

This project could be improved if we were to carry it out over a longer period of time. For instance, separation results would significantly improve if additional categorical factors, such as neighborhood, were included in the dimensionality reductions as dummy variables since it will be simpler to distinguish different room kinds and price ranges on the plot graphs. Nevertheless, the project's features gave insight into how specific variables affect the overall separation of data. Since LDA and MDS projections appear to make significant suggestions, further research into them is necessary. However, only PCA and LDA seem to offer a logical way to categorize rental properties if one were intending to do so in a basic manner. It's likely that by placing more focus on the limitations imposed by the separation of the specific data, rental properties should be offered that are distinctively different in terms of room type and cost.

VII REFERENCES

[1] https://www.kaggle.com/datasets/ivanchvez/ny-rental-properties-pricing

[2] https://www.mathworks.com/help/stats/

[3] https://www.sjsu.edu/faculty/guangliang.chen/Math250.html

[4] G. Chen, Introduction to Matrix-Based Data Science: Mathematics, Computing and Data, vol. 1. Springer Nature, 2023.