AirBnb Data Analysis of New York And Milano*

1st Ahmet Kaan Memioğlu
Engineering Department
Istanbul Kültür University
Istanbul, Turkey
1900005528@stu.iku.edu.tr

2nd Emrecan Üzüm
Engineering Department
Istanbul Kültür University
Istanbul, Turkey
1900005485@stu.iku.edu.tr

3rd Şükrü Erim Sinal
Engineering Department
Istanbul Kültür University
Istanbul, Turkey
1900003587@stu.iku.edu.tr

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

First of all we are using 2 dataset in our project. One for Milano and one dataset for New York city. these are for comparisons between each other in some criteria. We have chosen the New York and Milano datasets because they both provided a larger scale compared to other cities in Airbnb datasets and the data that those datasets provide are prominant.

II. Data Preprocessing

A. Imputation and Cleaning

In the features that we have imputed we either replaced the missing values by their calculated standard deviation or by their average mean so we could use the features that we have found suit of the percentage that they provide meaningfull data.

Also in those we have dropped they were either unnecessary or they did not give enough information about the status. Which means we didn't have enough statistical data for that feature to use in test forms or in visualization form. The cleaning process was quickly done with a drop function used in python below:

we have created two new dataframes after dropping the datasets. From there on the impution and cleaning was done in those dataframes using the same parameters so we could match the datasets. Also have replaced some features for example price, host-location, beds, host-review-scores and so on to give meaning for the missing values that are not given by default. We can change this by applying mean average imputing.

As for data analysing, we have defied simple outlier clearing function which uses features' quantile values to purge outliers. However, function has a downside of which has probability of shifting balance of features with ordinal values that has very limited range and constituting the majority

B. Labeling

We have used label encoding and ordinal encoding on categorical data which are for example: "room-type, property-type, host-has-profile-pic" and so on. By changing the data values of the features.

III. Data Visualization

In the visualization part we have first of all created a heatmap of all the contents that calculates the correlation of the features for the Milano and New York datasets.

After that we have made a pie chart for determining the popular Neighborhoods from the New York dataset. From the pie charts depiction we have made a visual type of bar chart that gets the first 10 popular neighborhoods and their frequencies for each dataset we have acquired.

To improve our accuracy at our machine learning algorithms we had to check the quality of our datasets. The quality was questionable so we went ahead and tried to determine the outliers for this process.

We have used boxplots for each New York and Milano. Because the New York is a greater and a larger populated and saturated city the quality compared to Milano was greater so the cleaning process didn't really require consideration but The Milano dataset was quite tricky to handle because there were a lot of outliers in the features we were trying to use on the Machine Learning algorithms.

We have used Review Scores, Availabilty30, Availabilty60, Availabilty90, Availabilty365 respectively for depicting them in boxplots. After the heatmaps and pie charts we have also correlation heatmap for the 4 host features we are using such as: 'hostlocation','hostresponserate','hostacceptancerate','hostresponsetime' and for property features such as: 'instantbookable','beds','bedrooms','accommodates','roomtype', 'price'. Same process was applied to the features we have used earlier. Outliers was getting ahead of the accuracies of Machine Learning algorithms and with careful consideration and deletion we have successfully improved accuracy and decreased the error rate of the Algorithms.

A. Heatmaps

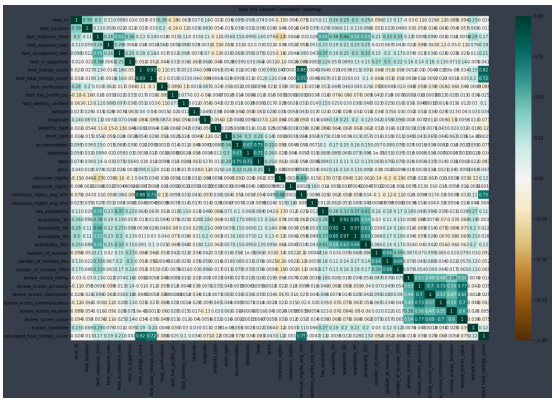

Fig. 1. New York Heatmap

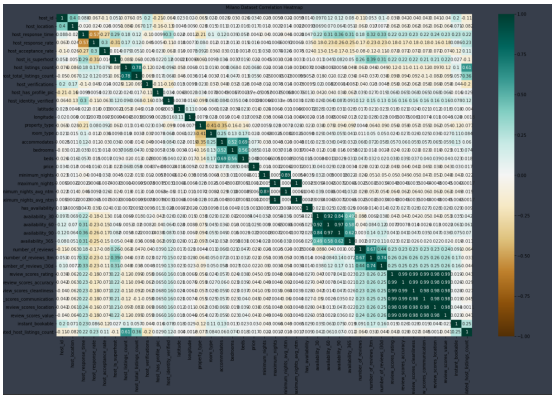

Fig. 2. New York Heatmap

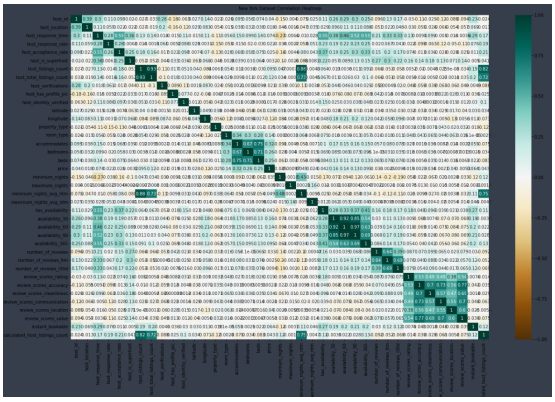

Fig. 3. New York Heatmap

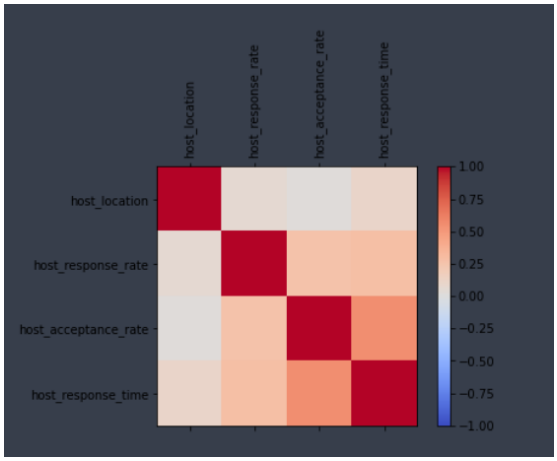

Fig. 4. Host features Heatmap

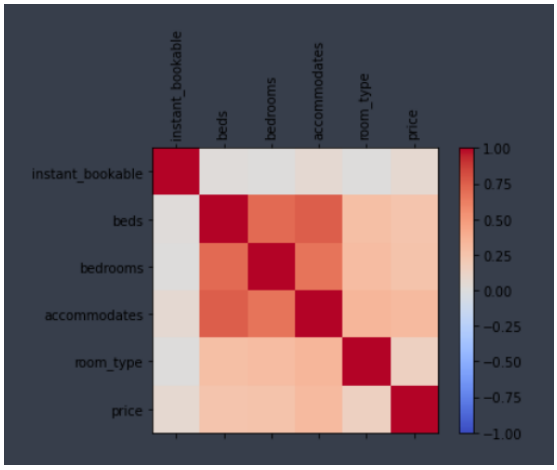

Fig. 5. Property Heatmap

B. Boxplots

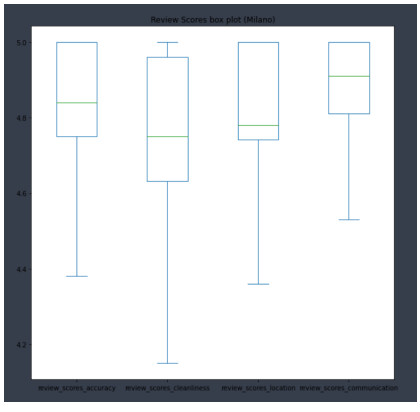

Fig. 6. Milano Boxplot 1

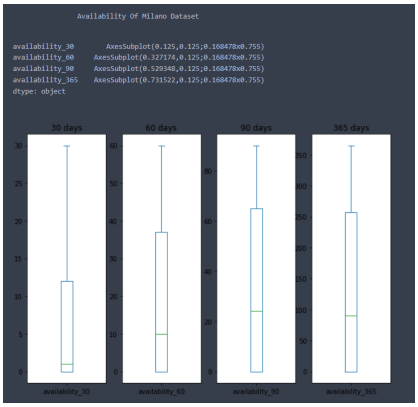

Fig. 7. Milano Boxplot 2

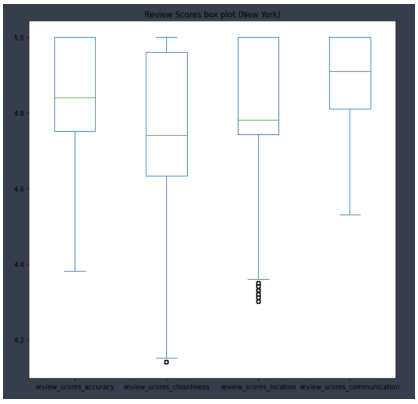

Fig. 8. New York Boxplot 1

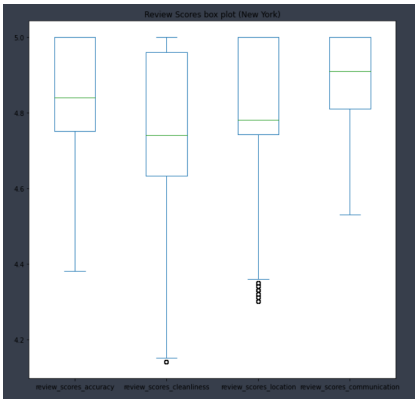

Fig. 9. New York Boxplot 2

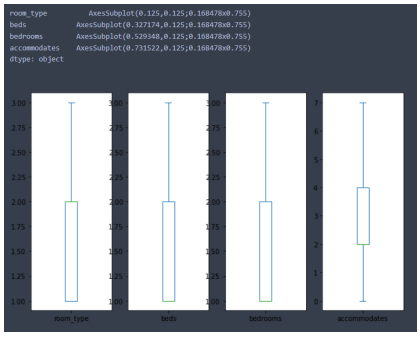

Fig. 10. Last Boxplot

C. Pie Chart and Bar Chart

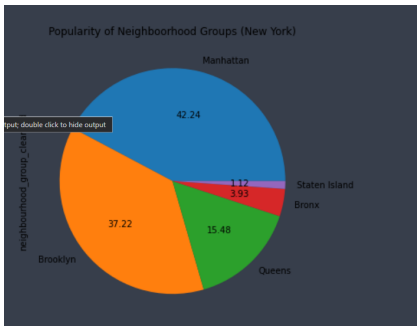

Fig. 11. Popularity Of Neighbourhood Pie Chart

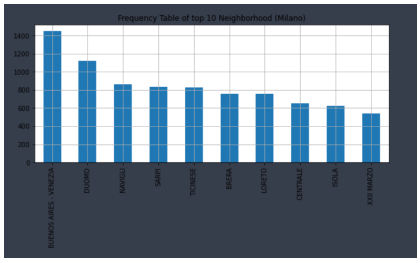

Fig. 12. Popularity Of Neighbourhood Bar Chart Milano

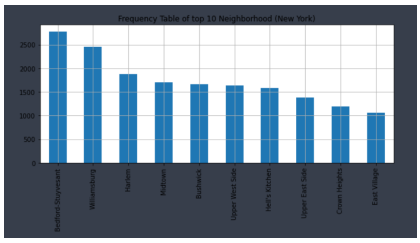

Fig. 13. Popularity Of Neighbourhood Bar Chart New York

STATISTICAL TEST

In the testing part we have decided to first of all check the normalization of the plotting charts that our features beholds.

Our normalization testing choice is the Anderson Darling test. This test essentially figures out whether a given sample of data is drawn from a given probability distribution.

In our testing we ought to look out for if Superhost features have some relation to another feature in our dataset so we saw fit this test in particular.

After the normality testing we moved onto Mann Whitney-U test. We chose this test to determine whether or not two samples of features are derived from the same population or not.

For this process we have divided our dataset into various divisions:"NewYorkNonSuperHost, MilanoSuperHost, MilanoNonSuperHost,NewYorkSuperHost" to test in 3 cases if the super host feature had any relations to :'hostAcceptancerate, hostlistingscount, price'.

MACHINE LEARNING

In the machine learning part, we have decided to make prediction on prices on both datasets. Firstly related features are used "room types","property type","accommodates","beds","bedrooms","instant bookable". And we used two continuous regression methods.
Afterwards, trained with all other features with outliers and non outlier versions in order to observe and increase test accuracy of model. Same operations are done for both datasets and results are as follows.

D. Price Prediction of New York

First Accuracy result we obtained is unacceptably lower than expected. Errors are also exceeded the acceptable ratio. As for KNN approach using 5k neighboor and "Manhattan

```
Linear-Regression Training Accuracy:0.28
Linear-Regression Testing Accuracy:0.28
Linear-Regression Mean Absolute Error: 0.003342090189901744
Linear-Regression Mean Squared Error: 2.0111539075901673e-05
Linear-Regression Root Mean Squared Error: 0.004484589064329269
```

Fig. 14. Lineer Regression Price Prediction of New York (Other features are excluded)

Distance" method, Accuracy goes bit lower than Lineer Regre. method and also error goes high a little. So we made cross validation sampling to observer whether any minor difference occurs.

```
KNN-Regression Training Accuracy:0.29
KNN-Regression Testing Accuracy:0.26
KNN-Regression Mean Absolute Error: 0.0032381942388705455
KNN-Regression Mean Squared Error: 2.0606935440793828e-05
KNN-Regression Root Mean Squared Error: 0.004539486252957908
```

Fig. 15. KNN Regression Price Prediction of New York (Other features are excluded)

The results we obtained from cross val., unfortunately couldn't make any significant difference in terms of accuracy. Therfore we make same operations with all the features included

```
Linear-Regression with Cross Validation Accuracy:0.27
Linear-Regression with Cross Validation Standart Deviation:0.08
KNN-Regression with Cross Validation Accuracy:0.28
KNN-Regression with Cross Validation Standart Deviation:0.08
```

Fig. 16. Lineer and KNN Reg. with Cross Val. Price Prediction of New York (Other features are excluded)

E. Price Prediction of New York with all Features

From what we have achieved so far, accuracy has increased noticiably from 0.28 to 0.50. Errors are also lowered to also 50 percent of previous model.

```
Linear-Regression Training Accuracy:0.49
Linear-Regression Testing Accuracy:0.47
Linear-Regression Mean Absolute Error: 0.0026222339096985225
Linear-Regression Mean Squared Error: 1.2705871750804797e-05
Linear-Regression Root Mean Squared Error: 0.0035645296675444853
```

Fig. 17. Lineer and KNN Reg. with Cross Val. Price Prediction of New York (Other features are included)

Same also can be said for KNN approach. Accuracy increased significantly and errors are reduced to almost 50 percent. Afterwards we implemented cross validation again to observe any probable changes

```
KNN-Regression Training Accuracy:0.48
KNN-Regression Testing Accuracy:0.41
KNN-Regression Mean Absolute Error: 0.002764357204247415
KNN-Regression Mean Squared Error: 1.4206487733527574e-05
KNN-Regression Root Mean Squared Error: 0.0037691494708392203
```

Fig. 18. KNN Regression Price Prediction of New York (Other features are included)

as it is understood that dataset itself not suprisingly gave lower accuracy. As far as data graphs are concerned, further we take random sample, lower accuracy we get. Even worse Milano dataset is far from being inefficient dataset.

```
Linear-Regression with Cross Validation Accuracy:0.36
Linear-Regression with Cross Validation Standart Deviation:0.39
KNN-Regression with Cross Validation Accuracy:0.32
KNN-Regression with Cross Validation Standart Deviation:0.32
```

Fig. 19. KNN Regression Price Prediction of New York (Other features are included)

F. Price Prediction of Milano

Situation in Milano is much different from York dataset. During data analysis we decided to use outlier cleared version directly in M.L. Since loss of feature was larger than expected and imputing would affect the model negatively.Despite we used the cleared Milano dataset, Accuracy of the model barely reaches ”0.10”, SQRE is extremely higher.

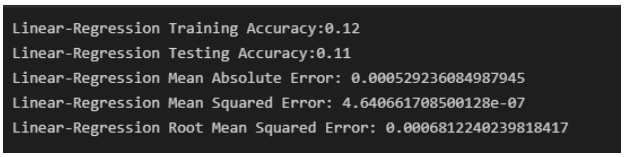

Fig. 20. KNN Regression Price Prediction of New York (Other features are included)

in K.N.N model was below 0 and resulted in negative accuracy. even if we did change the methods and neighbor,inefficaciousness of Milano dataset didn’t let us to improve model. As a result, we did step down from making model from Milano Dataset.

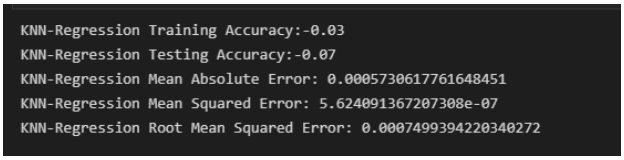

Fig. 21. KNN Regression Price Prediction of New York (Other features are included)