***Using Machine Learning to Understand Airbnb***

Clustering, Predicting Prices and comparing with long-term rental

Javier Rodríguez Roldán

Business Intelligence & Data Science Master´s degree. Dalarna University

Borlänge. Swedeen

[v23javro@du.se](mailto:v23javro@du.se)

*Abstract*— This paper aims to revise the literature to get an overview of how the short-term rental platform Airbnb has affected the tourism industry, the housing rental market and gentrification and also revise previous research about predicting the price of Airbnb. Furthermore, performing the machine learning algorithm of K-means clustering we found that Airbnbs can be categorized into various groups according to their characteristics and we can find different target customers associated with these Airbnb types. We also create a Random forest algorithm to predict the price of a new Airbnb listing achieving an R-squared score of 0.7666. We use this model to predict the price of a new listing in Barcelona to compare it with the equivalent house available for long-term rental to realize renting via Airbnb is much more profitable for the owner in these conditions. We also found the most important features for predicting Airbnb prices in order are: the number of nearby tourist attractions, how many people is the place suitable for and the distance to the city centre.

Keywords- Airbnb; Tourism; Housing Rental Market; Random Forest; K-means Clustering; prediction; housing rental comparison

# Introduction

Airbnb is a peer-to-peer accommodation platform. It was founded in San Francisco in 2008 by Brian Chesky, Joe Gebbia, and Nathan Blecharczyk [6]. It works as an intermediary between homeowners and short-term renters. What started as a start-up is now the benchmark in the short-term rental market and is now valued at over US$70 billion [1].

Since the arrival of short-term rental platforms such as Airbnb, there has been a very noticeable impact on tourism, the housing rental market and gentrification, affecting many people; some for better and some for worse.

The goal of this paper is to understand how Airbnbs can be separated to target different types of customers and to highlight the gap between the potential income a homeowner might have by renting via Airbnb or via traditional long-term rental in highly touristic European cities. We also create an algorithm to predict the price of new Airbnb listings.

It will be highlighted the considerable gap in terms of higher income Airbnb can offer to the homeowner and how this aggravates the problem of gentrification making it very difficult to mitigate and thus emphasising how important is that these platforms are well regulated by the state.

Contributing this way to further explore Airbnb´s impact on today´s society and as an alert for governments to create awareness that something should be done to find an equilibrium between tourists visiting their city and that affecting the local residents. The paper also has the intention for homeowners to evaluate whether or not it is worthwhile for them to choose one rental option over the other and categorize their house into one of the three groups or clusters to find out who might be the target customer and what price might be suited for the specifications of their house.

Regarding the fields and sectors Airbnb has had a bigger impact on we briefly introduce the three beforementioned ones:

## Tourism

The platform has become a very important player in the tourism sector. Travellers now have a new option to choose from as for where they stay when travelling that did not exist before.

The majority of the accommodations are located in large and medium-sized cities, especially in the city centre and tourist locations [3].

The transformation of regular housing into Airbnb rentals is strongly linked with nearby tourist attractions as well as supporting services. Findings from Barcelona [3] show that rather than spreading evenly throughout the city, Airbnbs are clustered in certain neighbourhoods.

## Housing Rental Market

Airbnb has also affected the housing market since it offers landlords an alternative opportunity; potentially affecting rental housing supply and affordability [7].

## Gentrification

Gentrification is the process in which residents from certain areas or neighbourhoods are forced to move out because of an increase in prices and property values caused by businesses or wealthy people investing in that area.

Literature shows various periods or “waves” of gentrification. The first three are caused by different factors such as government regulation or individuals and enterprises investing in the housing market [5]. The fourth wavem combines the intensified financialization of housing and the consolidation of pro-gentrification policies [8]. Lastly, the fifth wave[2] adds as a factor platform capitalism, especially Airbnb. Leading to price increases and ‘touristification’ of certain neighbourhoods, which results in house price inflation and more displacement.

# Literature Review

Previous research has been conducted on Airbnb pricing using machine learning algorithms, some with the goal of inferring to gain knowledge on how pricing works for Airbnb and some tried to predict it for new listings. It is worth mentioning that these studies have been conducted on different datasets about different regions of the world. Also, there is literature about how this platform has affected long-term rental such as in the case of Barcelona where there has been an increase in the rental prices due to Airbnb. The paper [4] concluded that for the average neighbourhood, Airbnb activity has increased rents by 1.9% and for neighbourhoods in the top decile of the distribution of Airbnb activity, rents are estimated to have increased by 7%.

An example of inference is [11] where the authors found that, even though prices are very heterogeneous, some trends can be found. Larger accommodations and a higher reputation have a positive impact on the price or Airbnb with many ratings are associated with lower prices. In general, it is seen that the more benefits (AC, elevator, TV, etc.) the higher the Airbnb is; it is also noticeable how the differences between countries and the situation of their economy and currency affect the price of the Airbnb too.

As for prediction, having tried different models, for the case of Airbnb listing in Boston [9], the best performance is achieved with extreme gradient boosting (XGBoost) with an RMSE of 59.9743 and an R-squared score of 0.632.

In the case of [12], XGBoost is also used as the prediction algorithm for New York data and the authors achieved an R-squared of 0.618.

Support Vector Regressor is the algorithm used in [10] for New York data as well and the authors achieve an R-squared of 0.690.

# Method description

## The Dataset

The analysed dataset consists of the Airbnbs´ characteristics of nine highly touristic European cities with 41,714 observations. The variables in the dataset that were used for the analysis can be seen in Table 1.

Table 1. Dataset variables

|  |  |
| --- | --- |
| **Variable** | **Meaning** |
| Price | Price in Euros by day of the Airbnb |
| Day | Whether it is a weekday or weekend |
| Room type | Entire apartment/home, private room or shared room |
| Person Capacity | How many people can stay in the Airbnb |
| Superhost | Whether the host is a superhost or not, which means that the host has had good reviews from previous customers |
| Multiple Rooms | Binary value measuring if there is more than one room or not |
| Business | If it has more than four offers |
| Cleanliness Rating | How the cleaning was evaluated by previous customers |
| Guest Satisfaction | How the satisfaction was evaluated by previous customers |
| Bedrooms | How many bedrooms it has |
| City Center (km) | Distance in kilometres to the city centre station |
| Metro Distance (km) | Distance in kilometres to the closest metro station |
| Attraction Index | Number of tourist attractions nearby |
| Restaurant Index | Number of restaurants nearby |

Cleaning the data, we removed some outliers, having in mind that “Price” is the variable to predict or dependent variable and it does not follow a Gaussian distribution, as can be seen in Figure 1, we used the Inter Quartile Range technique to establish a top threshold which is €527.41 per night, Airbnbs with higher prices were removed as they are considered outliers; they could be analyzed separately, but for this analysis, they would have negatively affected the results.

We also decided to remove the following variables that are not included either in Table 1: "Normalised Restaurant Index", "Normalised Attraction Index", "Shared Room" and "Private Room". Since they were already represented in some other variables ("Restaurant Index", "Attraction Index", “Room Type”) and we would have had duplicated data in case of maintaining them. In the case of normalized variables, we prefer the raw data to pre-process it how we saw fit.

For the case of clustering, the pre-processing involved keeping the numerical variables only and scaling the data by removing the mean and dividing the standard deviation for each variable. This is done so all variables have the same impact when calculating the (in this case) Euclidean distances and the clustering is not led by some variables more than others.

In the case of random forest, the pre-processing nivolved creating dummy variables for the categorical ones so that Python could process them.

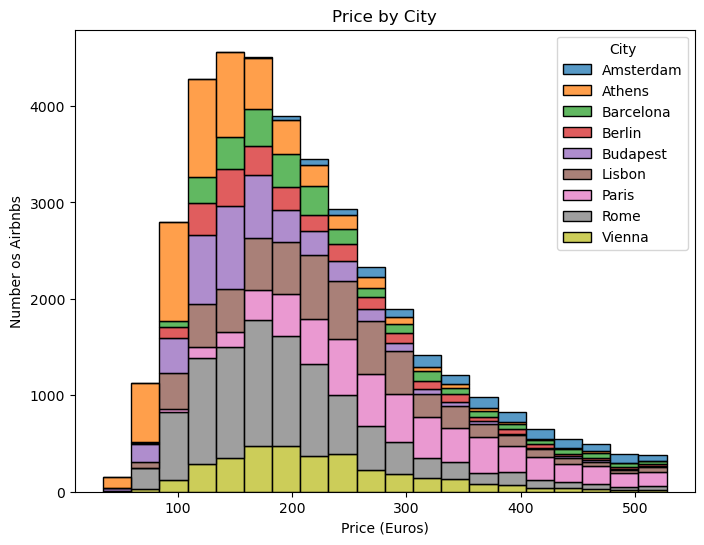


Figure 1. Price distribution for each city

From Figure 1 we can see that more than 50% of the Airbnbs have a price between €110 and €280 per night and the city with more representation in the data is Rome.

## Data Mining Methods

The used models are K-means clustering and random forest.

Random forest was used as a feature selection method as well as for predicting the price of a new listing. The most important variable found is the Attraction Index, closely followed by whether or not the Airbnb is located in Paris, although this second one is not relevant for this analysis, we kept it because the performance of the model is better, but the idea is to generalize to all touristic European cities. The detailed results can be seen in Figure 2. This was done for possible future research and to perform another random forest only using the most relevant variables.

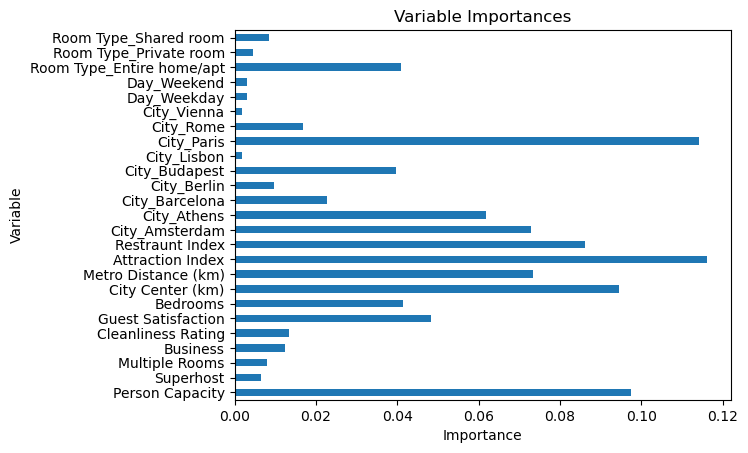


Figure 2. Most important variables for predicting Airbnb prices

In Figure 2 we can see that the most important variable is the one measuring the number of touristic attractions near the Airbnb, followed by Person Capacity and Distance to the City Centre. That ignoring City\_Paris as we mentioned in the previous paragraph to generalize to any city.

The importance, in this case, is measured by calculating the average impurity decrease caused by including each variable on the built trees.

It is worth mentioning that building the random forest algorithm with only the three most important variables gives a considerably worse performance (0.243 R-squared less) than including all of them. The reason behind this is that random forest is an ensemble model which is based on combining a lot of weak learners, so even if those variables are not great at predicting the price, in combination with all the others they do make a much better prediction model.

The goal of K-means clustering is to understand the types of Airbnb there are according to their specifications and for homeowners to understand which one they have or potentially will have and maybe target different customers or tourists who might be interested in their place to make better decisions about pricing or how to promote the listing on Airbnb´s website.

To do this, we performed the algorithm and calculated each cluster’s centroid. The centroid is the average of all points which correspond to the cluster and it is the point that best represents it. To get to that point the algorithm is calculating as many random centroids as clusters we indicate, then it calculates the distance from all data points to all centroids and assigns each point to the closest centroid, after this it recalculates the centroid of each cluster based on the average of all the values that were assigned to it. Now some points will be re-assigned to another cluster so new centroids will be calculated and the same process applies until the centroids do not change anymore, which are the ones that minimize the distance from the points to their centroid and in between points of the same cluster.

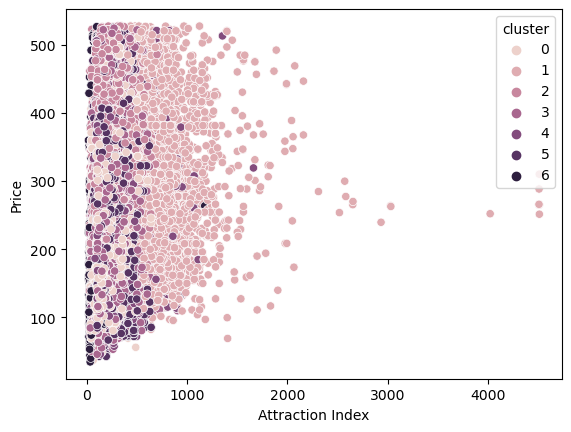


Figure 3. Clusters and centroids representation

The specific results will be discussed in Table 2, but in Figure 3 it can be appreciated that there are seven clusters. The Y axes is the Price since it is the variable we want to focus on and the X axes is the Attraction Index because it is the most relevant one for predicting the price according to the random forest model. The decision to select seven clusters can be contrasted in the following elbow plot in Figure 4.

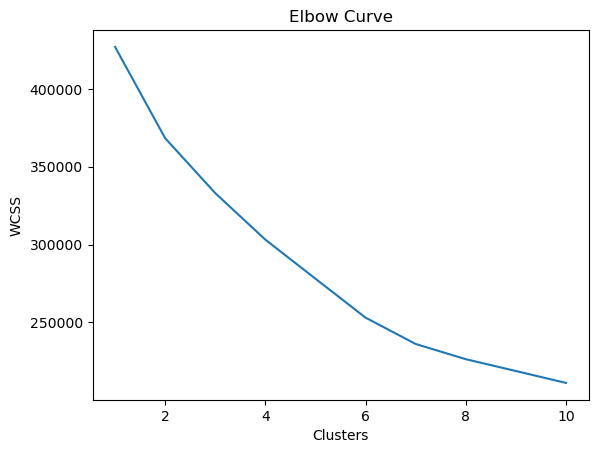


Figure 4. Elbow curve plot to evaluate how many clusters were needed

From Figure 4 we see that the Within-Cluster Sum of Square (WCSS) starts flattening after seven clusters, that is how we decide that is the optimal number of clusters in this case. This decision is not deterministic and it could have been another number and would not necessarily be wrong. The main disadvantage of this method is, in fact, that the number of clusters has to be decided by the user, no rule or algorithm decides for us, and that decision has a great impact on the results.

The random forest algorithm provides us with a tool for predicting the price of a new Airbnb listing given the house specifications. This would be a very useful tool for a homeowner evaluating whether or not to start with Airbnb or switch to Airbnb instead of the long-term rental by predicting the price that people would be willing to pay according to the already existing supply.

The predictions are achieved by creating many decision trees, these are created by choosing a random set of variables on a bootstrapped dataset, which is a randomly picked set of observations with replacement, so an observation can be present once, repeated more than once or not be present at all in the bootstrapped dataset. Once all trees are built all of them predict the price of the Airbnb and the average result is the output of the random forest.

# Results and Analysis

The performance of the random forest model was measured using the Mean Squared Error (MSE) and the R-squared. The result MSE is 2,388.88, (48.88 RMSE) and 0.7666 respectively, which is better than the ones reviewed in the literature. Meaning 76.66% of the variation in Price is explained by the independent variables included in our model.

As mentioned before the main reason to do clustering is to find what specific characteristics each Airbnb cluster has. In Table 2 we can see the values for all variables in each centroid´s cluster.

Table 2. Centroid values for each cluster

|  |  |
| --- | --- |
|  | |
|  | **Clusters** | | | | | | | | |
| **0** | | **1** | **2** | **3** | **4** | **5** | **6** |
| **Price (Euros)** | | 204,46 | | 287,23 | 263,76 | 185,03 | 214,44 | 195,45 | 185,99 |
| **Person Capacity** | | 2,67 | | 3,26 | 5,00 | 2,72 | 3,00 | 2,92 | 3,07 |
| **Superhost** | | 0,30 | | 0,27 | 0,33 | 0,36 | 0,02 | 0,24 | 0,28 |
| **Multiple Rooms** | | 0,00 | | 0,29 | 0,23 | 1,00 | 0,18 | 0,00 | 0,35 |
| **Business** | | 0,00 | | 0,43 | 0,34 | 0,00 | 0,54 | 1,00 | 0,19 |
| **Cleanliness Rating** | | 9,62 | | 9,47 | 9,62 | 9,59 | 7,11 | 9,54 | 9,56 |
| **Guest Satisfaction** | | 95,59 | | 93,03 | 94,78 | 94,43 | 71,25 | 93,02 | 94,25 |
| **Bedrooms** | | 0,90 | | 1,10 | 2,13 | 0,97 | 1,01 | 0,88 | 1,19 |
| **City Center (km)** | | 3,06 | | 1,76 | 2,41 | 2,72 | 2,62 | 2,12 | 7,79 |
| **Metro Distance (km)** | | 0,50 | | 0,67 | 0,52 | 0,52 | 0,48 | 0,44 | 3,25 |
| **Attraction Index** | | 213,70 | | 757,00 | 218,45 | 241,04 | 268,00 | 243,24 | 97,42 |
| **Restaurant Index** | | 436,59 | | 1687,41 | 466,94 | 520,14 | 589,80 | 537,48 | 187,79 |

We can see in Table 2 that all clusters have some differentiation. Cluster 0 is far from the city centre but has the highest cleanliness rating tied with cluster 2 and the best guest satisfaction, it is priced in the middle and it has just one room and capacity for few people, it also does not have many restaurants or attractions nearby; cluster 1is the most expensive one, it is farther to the metro station but it has the highest number of restaurants and touristic places nearby and is the one closest to the city centre; cluster 2 is the one where most number of people can stay and the highest number of rooms, it is on the higher price range ,it is on the middle location-wise and great cleanliness rating; cluster 3 is the lowest price, although it does have more than one room, it is able to fit few people and it does not have more than four offers but it is on the middle as for location; cluster 4 is the one with lowest Superhost rank, Guest Satisfaction and Cleanliness Rating mark, although it is near touristic attractions and restaurants and closer the metro station; cluster 5 only has one room but it has more than four offers, it is also the closest to the metro station, it is close to the city centre and it is priced on the lower end; lastly, cluster 6 is the one with the least number of restaurant and attractions nearby, it is the farthest one to the city centre and metro station, but the price is very close to being the lowest one and the person capacity and number of rooms is on the higher end.

We can see how some listings are better suited for certain types of customers or tourist. We have some better suited for people who prefer multiple rooms, others are suited for people with a higher budget that prefer to stay in a better location, some people are willing to pay more if the reviews from previous customers have been good and some might prefer being a bit further to the centre, but closer to a metro station.

It is useful to categorize clusters in order to see what price is fair compared to the existing offers available and try to find out what is the target customer. For instance, the low-cost Airbnb, clusters 3 and 6, although they are not in a great location, they might be a good option for people on a limited budget, such as students and younger people; cluster 2 is appropriate for larger groups which could be a group of friends of families with three or more children; clusters 0 or 5, which have a better location but just one room might be perfect for couples; cluster 4 could be for those who prefer to save some money even wanting a good location by staying at a place with worse reviews by previous customers; or cluster 1 which is close to the centre might be well-suited for customers who book the place for fewer days and maybe there can be more rotation in those cases since they will need less time to see the main points of the city.

After this analysis, we proceeded to predict the price for an Airbnb in Barcelona for two people with one bedroom and the average of all the other features. The resulting price is €327.61 per night. We then searched the most famous rental website in Spain to look for an apartment with the same characteristics to see at what price it was being offered and compare. It is very hard to compare both since for the rental market there are very important aspects that affect the price significantly and for Airbnb are not relevant and vice versa, such as how many years since the construction or last reform, the number of square meters, etc. The average rental price is around €1,500 a month, we insist this is an estimate and could vary greatly in both directions, more or less price. That price is linked with a great location, having many tourist attractions and restaurants and also being close to the centre and metro station, around 50m², one bedroom, third or higher floor and the apartment is in good condition.

Comparing the two options, short-term rental with Airbnb and long-term traditional rental, it is clear that Airbnb is more profitable. Obviously, there are many factors for a particular Airbnb to be in demand and it will most likely not be booked every day, but just taking weekends alone the revenue exceeds the one of the long-term rental by €1,012 (314 \* 8 weekend days a month – 1,500 = 2,512 – 1,500) or it can also be viewed as earning 60% of the amount you could earn in Airbnb. Although from this amount you would probably have to deduct an amount for cleaning services and, depending on how the owner of the Airbnb advertises it, up to 16% commission paid to Airbnb, in that case, the extra revenue would be not that high (€850 profit for the case of not paying for professional cleaning services, but the owner taking over all fees). This is only in the scenario where the Airbnb was booked just for the weekends, which in the cities that have been analyzed is not usual, they receive a lot of tourists and the demand for accommodation, including Airbnbs is quite high throughout the year and especially high during certain times such as summer.

# conclusion

Airbnb has had positive and negative impacts on today’s society and economy.

K-means clustering analysis allowed us to distinguish different types of Airbnb according to their characteristics and reviews from previous customers and they might be better suited for different types of tourists.

Using random forest we achieve a prediction model with a MSE score of 2,388.88. We also found the most important variables to predict the price are: the number of tourist attractions close to the Airbnb, the Person Capacity and the distance to the City Centre.

A homeowner in the centre of Barcelona could earn more by renting through Airbnb than traditional rental. The first option is €314 per night and the traditional rental is around €1,500 per month. Even considering the higher cost of maintaining an Airbnb and the variability of bookings one can achieve, with not many booked days it matches the rental option. Keep in mind that these are very challenging issues to compare.

A key takeaway from this analysis is that it is understandable why homeowners in the centres of cities that receive a lot of tourism might choose Airbnb instead of traditional rental. For local residents this has a negative impact as it aggravates gentrification.

This research and analysis could be further improved. As the opposition mentioned, it could be interesting to see differences across cities for example. There can be a better and more fair comparison between both types of rental methods too and much more research can be done related to Airbnb.

##### References

1. (n.d.). Airbnb, Inc. (ABNB). Yahoo Finance. Retrieved April 23, 2023, from https://finance.yahoo.com/quote/ABNB/
2. Aalbers, M.B. (2019), Introduction To The Forum: From Third To Fifth-Wave Gentrification. Tijds. voor econ. en Soc. Geog., 110: 1-11. https://doi-org.www.bibproxy.du.se/10.1111/tesg.12332J.
3. Adamiak, C., Szyda, B., Dubownik, A., & García-Álvarez, D. (2019). Airbnb Offer in Spain—Spatial Analysis of the Pattern and Determinants of Its Distribution. ISPRS International Journal of Geo-Information, 8(3), 155. MDPI AG. Retrieved from <http://dx.doi.org/10.3390/ijgi8030155>
4. Garcia-López, Miquel-Àngel & Jofre-Monseny, Jordi & Martínez-Mazza, Rodrigo & Segú, Mariona. Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona, Journal of Urban Economics, Volume 119, 2020, 103278, ISSN 0094-1190, https://doi.org/10.1016/j.jue.2020.103278
5. Hackworth, J. and Smith, N. (2001), The changing state of gentrification. Tijdschrift voor economische en sociale geografie, 92: 464-477. <https://doi-org.www.bibproxy.du.se/10.1111/1467-9663.00172>
6. Hoffman, L. M., & Schmitter, H. B. (2020). Airbnb, short-term rentals, and the future of housing. Taylor & Francis Group
7. Hui Li, Yijin Kim, Kannan Srinivasan (2022). Market Shifts in the Sharing Economy: The Impact of Airbnb on Housing Rentals. Management Science 68(11):8015-8044. https://doi.org/10.1287/mnsc.2021.4288
8. Lees, L., T. Slater & E. Wyly (2008), Gentrification. New York: Routledge
9. Liu, Yinyihong. (2021) Airbnb Pricing Based on Statistical Machine Learning Models. https://www.researchgate.net/publication/370698841\_Predicting\_Airbnb\_Listing\_Price\_with\_Different\_models
10. Rezazadeh Kalehbasti P, Nikolenko L, Rezaei H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. Lecture Notes in Computer Science.:173–184. doi:https://doi.org/10.1007/978-3-030-84060-0\_11
11. Troy Lorde, Jadon Jacob, Quinn Weekes. (2019). Price-setting behavior in a tourism sharing economy accommodation market: A hedonic price analysis of AirBnB hosts in the caribbean. Tourism Management Perspectives. Volume 30.Pages 251-261.ISSN 2211-9736. https://doi.org/10.1016/j.tmp.2019.03.006
12. Zhu, Ang, Li, Rong & Xie, Zehao. (2020). Machine Learning Prediction of New York Airbnb Prices. <https://ieeexplore-ieee-org.www.bibproxy.du.se/stamp/stamp.jsp?tp=&arnumber=9253078&tag=1>