

Forecasting AirBnb revenues in Geneva

The context and the problem

- Airbnb is a popular website to offer short time accomodation to customers
- A municipality might want to know the situaition in its city to better understand if Airbnb is going well or not in its territory and what are the main factors affecting the revenues for such service
- Some renter might also want to have insights on how to improve the revenues from its rented properties

The problem:

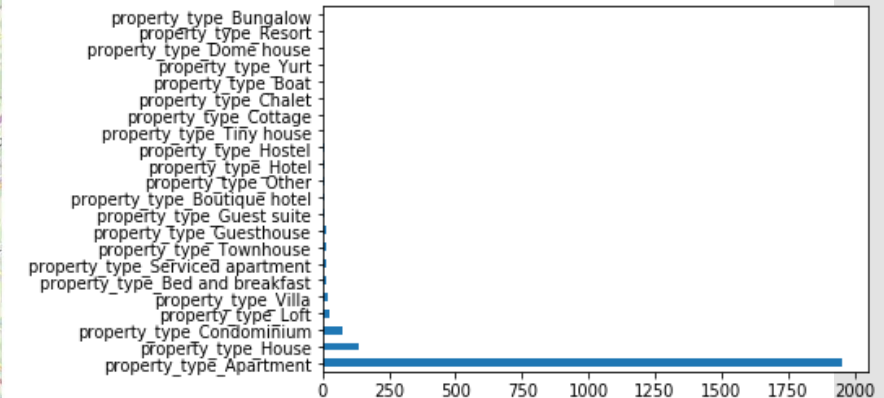
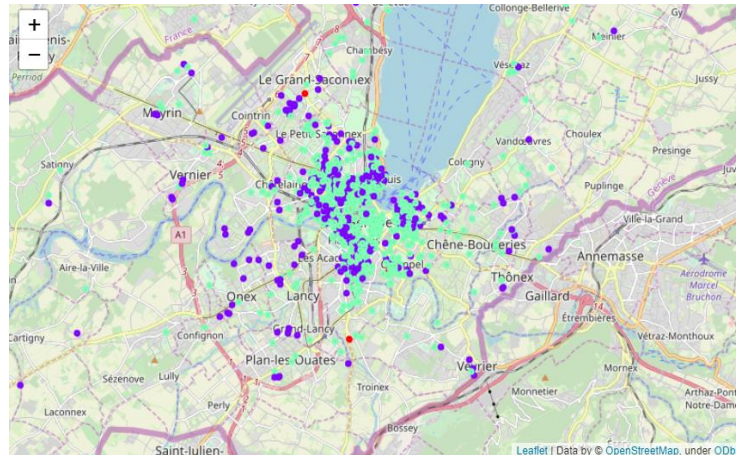
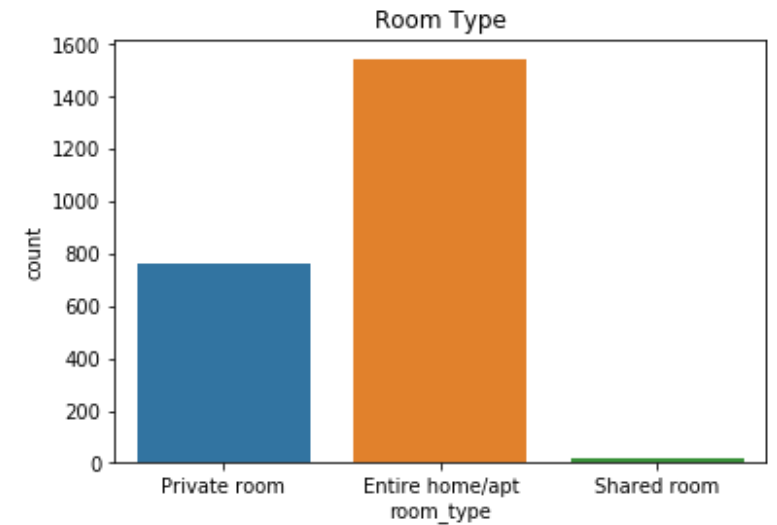
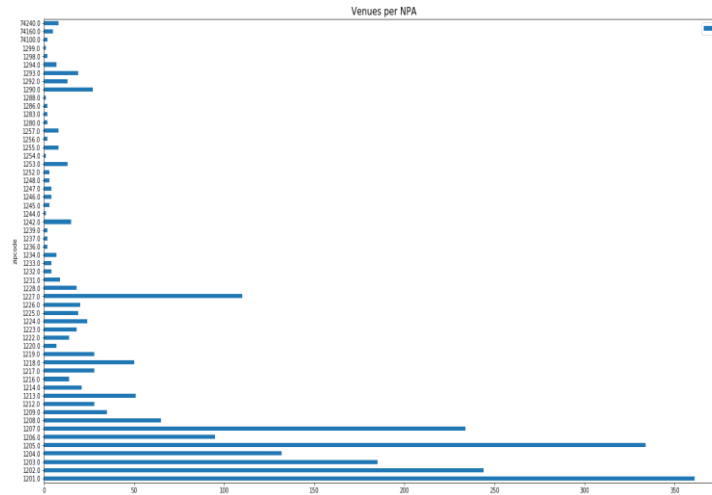
- In the project we foreacast prices and annual revenues of homes and appartments in Geneva, Switzerland, using data from
 - <http://insideairbnb.com/get-the-data.html>
 - Foursquare API
 - Geopy

Exploratory analysis

Most venues rented are entire homes or apartments.

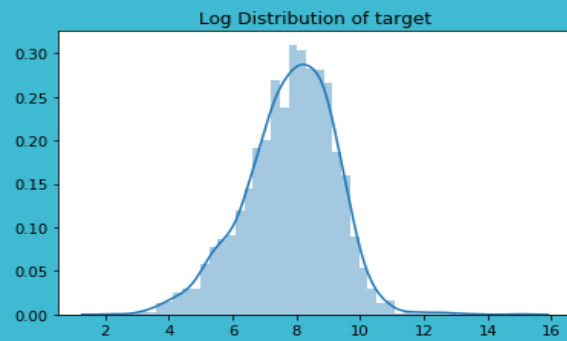
Most venues are typically rented in apartments and not in houses.

The houses and apartments are more or less evenly distributed in the city center while the number is decreasing in more peripheral neighborhoods.

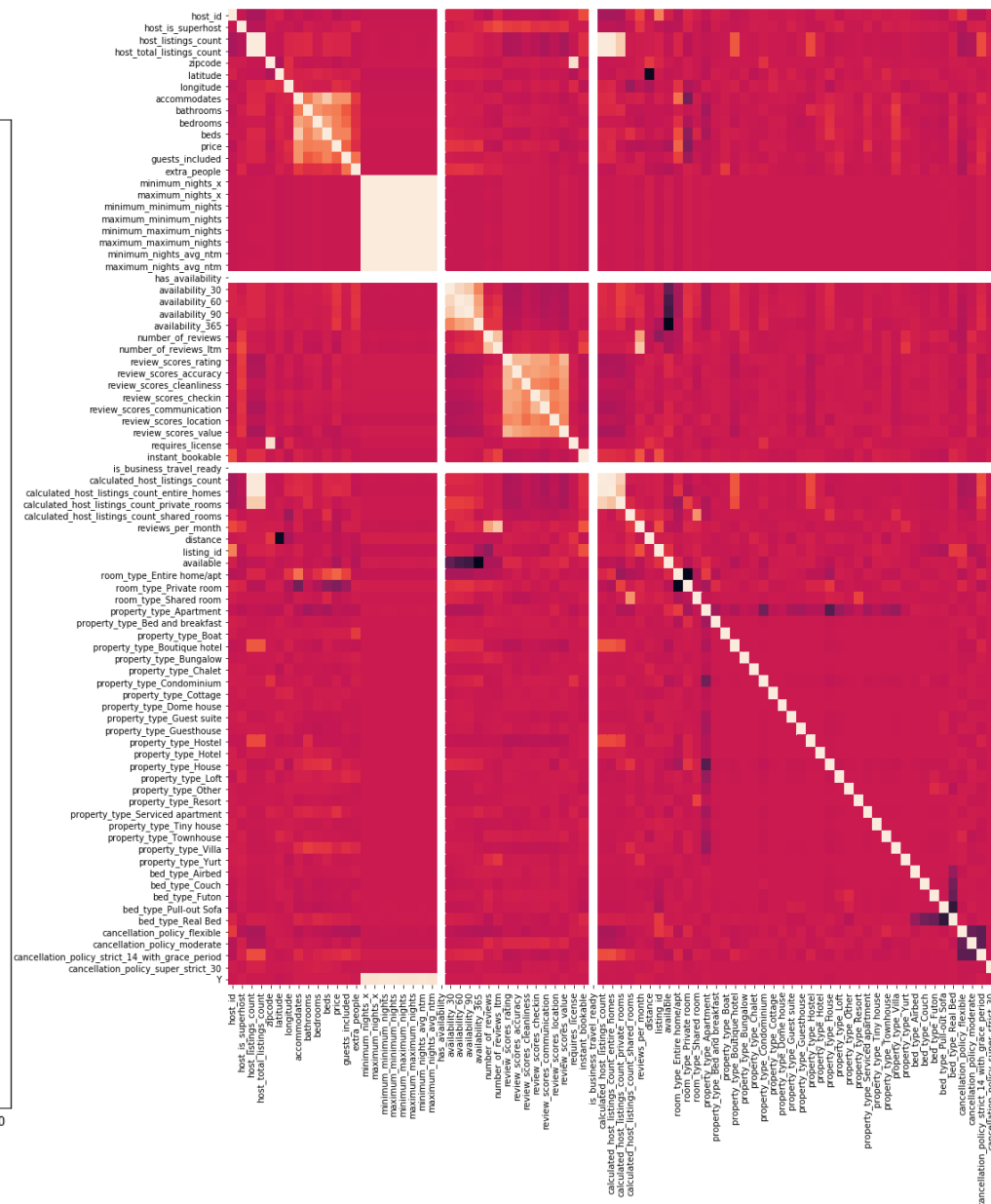


We apply a logarithmic transformation to the target variable to make its distribution less skewed towards high values

Few parameters seem to correlate heavily with the target.



Feature	Importance
is_business_travel_ready	0.000000
has_availability	0.000000
minimum_nights_avg_ntm	0.000000
maximum_minimum_nights	0.000000
minimum_minimum_nights	0.000000
minimum_nights_x	0.000000
minimum_maximum_nights	0.000000
maximum_nights_avg_ntm	0.000000
maximum_nights_x	0.000000
maximum_maximum_nights	0.000000
reviews_per_month	0.000000
number_of_reviews	0.000000
distance	0.000000
guests_included	0.000000
price	0.000000
cancellation_policy_strict_14_with_grace_period	0.000000
accommodates	0.000000
review_scores_cleanliness	0.000000
beds	0.000000
bedrooms	0.000000
review_scores_rating	0.000000
number_of_reviews_ltm	0.000000
review_scores_accuracy	0.000000
review_scores_value	0.000000
review_scores_checkin	0.000000
bathrooms	0.000000
host_is_superhost	0.000000
property_type_Apartment	0.000000
review_scores_communication	0.000000
availability_30	0.000000
availability_60	0.000000
bed_type_Real_Bed	0.000000
property_type_Serviced_apartment	0.000000
property_type_Villa	0.000000
extra_people	0.000000
availability_90	0.000000
property_type_Dome_house	0.000000
room_type_Entire_home/apartment	0.000000
available	0.000000
property_type_Yurt	0.000000
instant_bookable	0.000000
property_type_Bungalow	0.000000
property_type_Townhouse	0.000000
zipcode	0.000000
room_type_Private_room	0.000000
cancellation_policy_super_strict_30	0.000000
property_type_Boat	0.000000
property_type_Hotel	0.000000
availability_365	0.000000
property_type_Cottage	0.000000
requires_license	0.000000
property_type_Resort	0.000000
property_type Loft	0.000000
bed_type_Airbed	0.000000
property_type_Other	0.000000
property_type_Guest_suite	0.000000
property_type Chalet	0.000000
property_type_Bed_and_breakfast	0.000000
property_type_Tiny_house	0.000000
bed_type_Couch	0.000000
bed_type_Futon	0.000000
property_type_Guesthouse	0.000000
calculated_host_listings_count_entire_homes	0.000000
property_type_Hostel	0.000000
host_listings_count	0.000000
host_total_listings_count	0.000000
bed_type Pull-out_Sofa	0.000000
property_type_Boutique_hotel	0.000000
calculated_host_listings_count	0.000000
calculated_host_listings_count_shared_rooms	0.000000
room_type_Shared_room	0.000000
host_id	0.000000
review_scores_location	0.000000
longitude	0.000000
property_type Condominium	0.000000
calculated_host_listings_count_private_rooms	0.000000
cancellation_policy_moderate	0.000000
listing_id	0.000000
property_type_House	0.000000
cancellation_policy_Flexible	0.000000
latitude	0.000000



Modelling

The modeling pipeline is tuned and validated on a test set.

The results are shown in the features. As suggested by the correlation matrix the target variable, annual revenues per year is pretty well correlated with number of days the house is rented and the number of reviews per month. Actually this is the biggest result of the model.

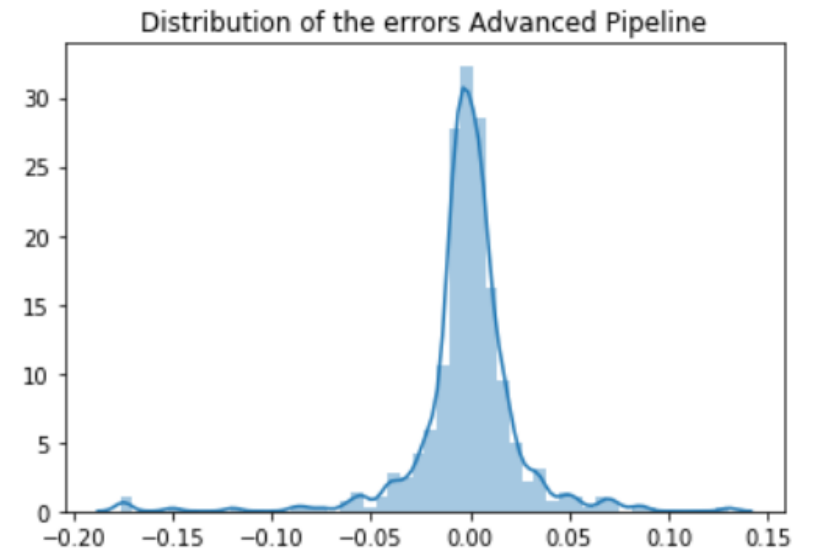
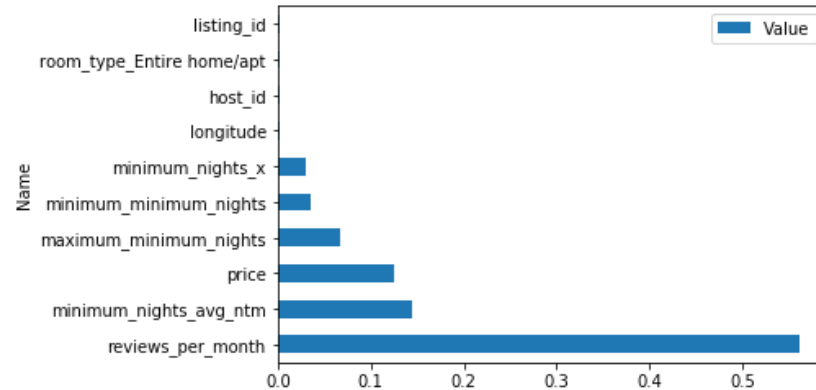
Target Variable = Annual revenues per venue

Model: RandomForestRegressor

CV: kFold

TestTrain: 20%/80%

Average error on annual revenues : 11.7%

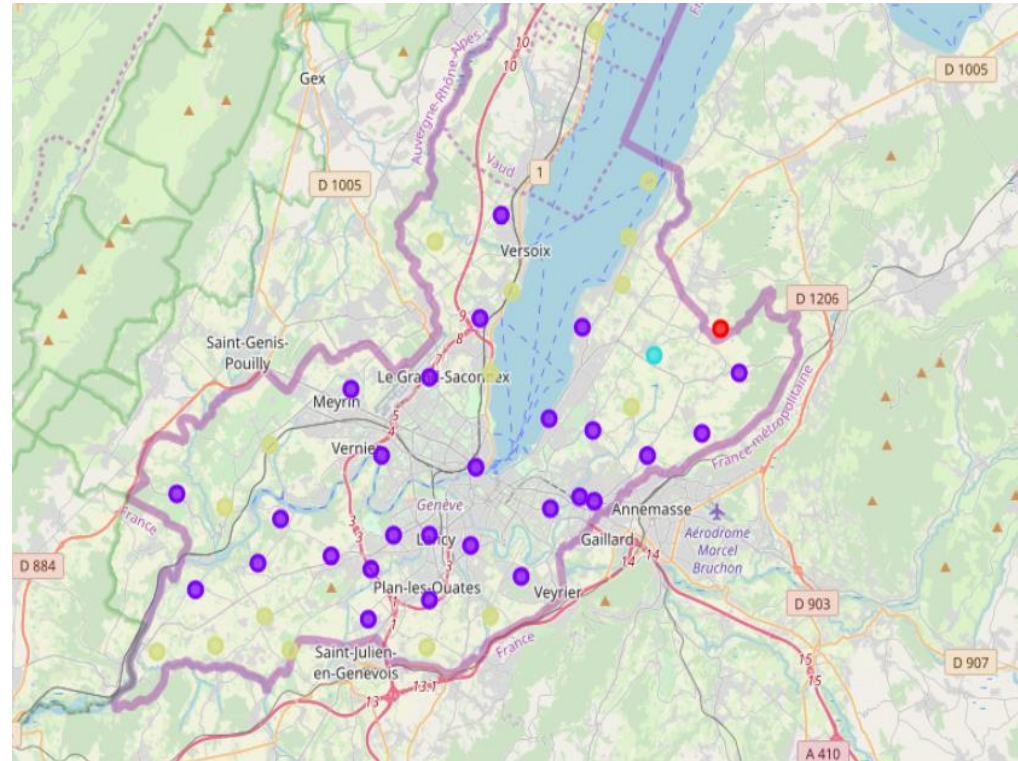


K-means clusters

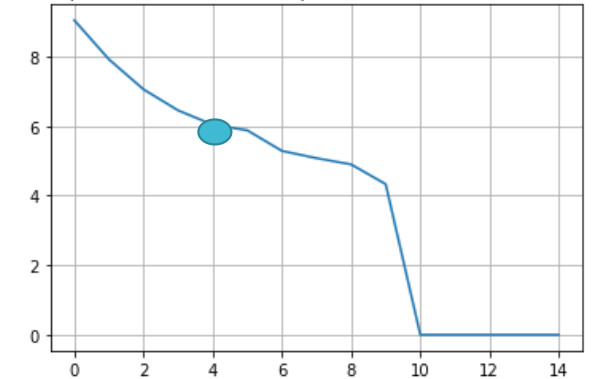
To improve quality of the model the foursquare api is used to make cluster between the neighborhoods in geneva. The results are shown in the figures.

The model does not get a better accuracy using this approach

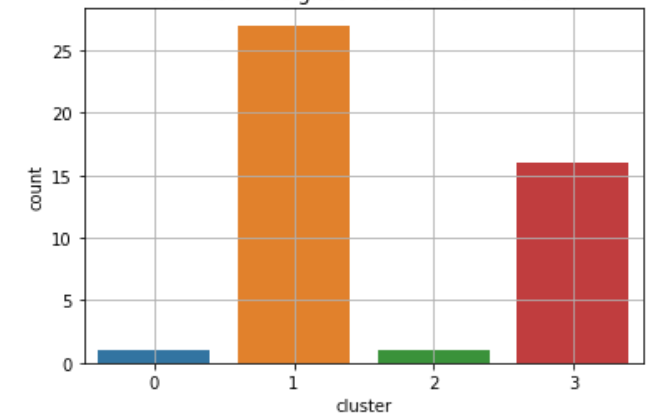
Target Variable = Annual revenues per venue



Sum of squared distances of samples from their closest cluster center

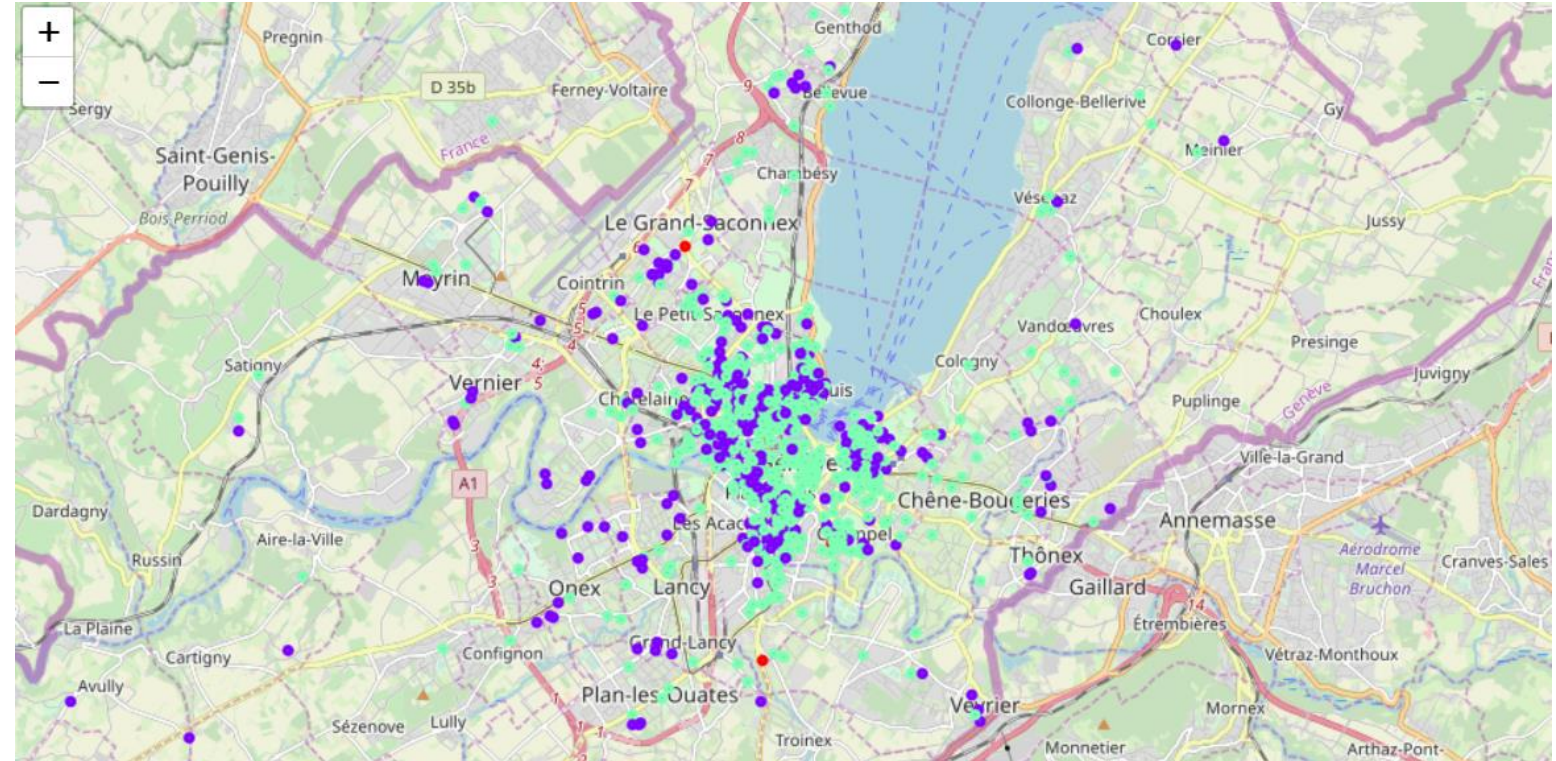


Number of neighborhoods in each cluster



Average error on annual revenues : 12.7%

The solution



- A random Forest Regressor is trained on a sample to predict the revenues for a given house the accuracy is computed as the MSE of 11%
- The most important parameter is found to be the number of reviews an apartment has got.
- Using neighborhoods characteristics to reinforce the learning has no effect. This might be to the fact that Geneva is a relatively small city with lot of similarities in its territory.
- Municipalities and renters can have a better look at the scenario of AirBnb thanks to this work and might try to understand if their property can get more value from AirBnb or from other renting strategies.

Patrizio Canzi