Forecasting AirBnb prices in Geneva

Author: Patrizio Canzi

Introduction

Airbnb is a popular website to offer short time accomodation to customers and in the recent years it has been used in a more and more massive way by users all around the world.

A municipality might want to know the situaiton 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 :

An analytic tool to provide insights on AirBnb usage in a city might be interesting for the municipality and to get to know the city caracteristics from the point of view of AirBnB users.

Oppositely a renter might want to monitor its rented properties and see if he is investing properly its assets.

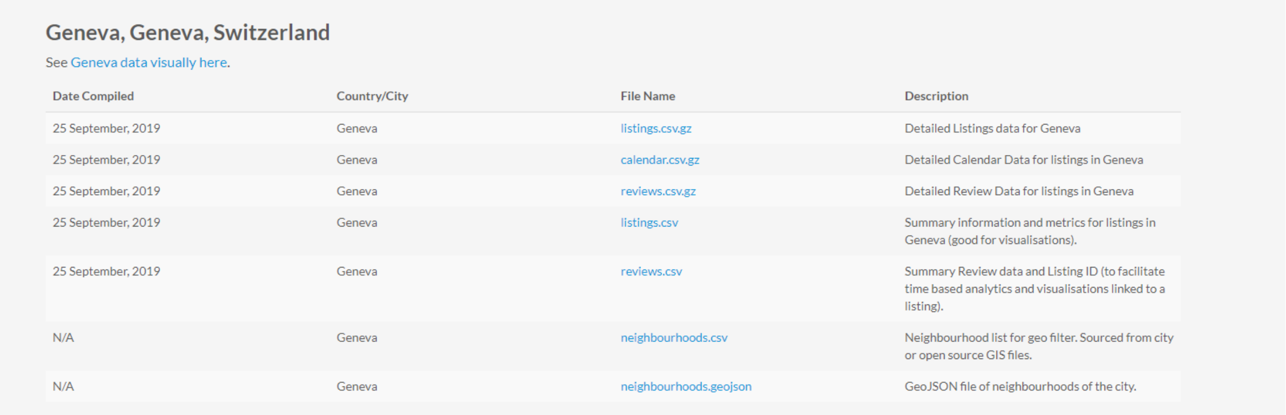
In both cases it is crucial to be able to forecast accurately the amount of money you can make out of a property in AirBnb

Data

Inside AirBnb

In the project we foreacast prices and annual revenues of homes and appartments in Geneva, Switzerland, using data from Inside AirBnb which puts at disposal the data from airBnbhttp://insideairbnb.com/get-the-data.html.

In this database you will have acces to a lot of interestin features from the listings the reviews and the calendar from AirBnb



Geneva, Geneva, Switzerland

See Geneva data visually here.

Date Compiled	Country/City	File Name	Description
25 September, 2019	Geneva	listings.csv.gz	Detailed Listings data for Geneva
25 September, 2019	Geneva	calendar.csv.gz	Detailed Calendar Data for listings in Geneva
25 September, 2019	Geneva	reviews.csv.gz	Detailed Review Data for listings in Geneva
25 September, 2019	Geneva	listings.csv	Summary information and metrics for listings in Geneva (good for visualisations).
25 September, 2019	Geneva	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Geneva	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	Geneva	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

Foursquare API

the foursquare api is used to enrich the different neighnborhoods of geneva to assess if the composition of the neighbohoods shops and facilities has an interesting impact on the AirBnb price.

Geopy Nominatim

Geopy is used to enrich the neighborhoods points and adresses to get some geospatial data from OpenStreetMap. It has to be noticed that in the Inside airBnb you also have some GeoJson data about neighborhoods but the fact that they are in SwissCoordinates complexifies their handling and therefore Geopy is preferred. For more information https://geopy.readthedocs.io/en/stable/

Methodology

To compute annual revenues from AirBnb listings a rule to compute the annual revenues might be found as the database we have does not have this data but we will have to build it.

The idea on computing the work is stated in the site and reported here accordingly credits to the authors :

Inside Airbnb

A conservative occupancy model has been built in order to estimate **Occupancy Rates**, **Income per Month** and **Nights per Year**. More information on the methodolgy of the occupancy model can be found in the disclaimers.

Inside Airbnb: Geneva uses the following parameters:

- A **high availability** metric and filter of **60 days per year**
- A **frequently rented** filter of **60 days per year**
- A **review rate** of **50%** for the number of guests making a booking who leave a review
- An average booking of **3 nights** unless a higher minimum nights is configured for a listing.
- A maximum **occupancy rate** of **70%** to ensure the occupancy model does not produce artifically high results based on the available data

If you are a data scientist, urban or public policy planner, researcher or journalist, get the data, analyze and publish your results.

Acknowledgements

- Thank you to Charles-André Aymon, a journalist working with Bilan, for requesting the data and for assistance with geographic, social and legal context for Geneva.

The original dataframe has 3200 entries and we apply a custom cleaning pipeline.

The cleaning methodology has succesive steps to get to a good dataframe

1) Drop all columns whith more than 750 empty values
2) Drop all null values by row.

The final result is about 2300 lines.

The most important categories in the listing dataframe are mainly related to :

- Caracteristics of the owner (photo, name, url, number of rented units)
- Dwelling (type of building, room, position in the city, bathroom type, price per night,)
- Duration of the rent (min and max durations, and availabilities)
- Reviews (mark of reviews in different categories, number of reviews)

An external table with detailed reviews is also availabe.

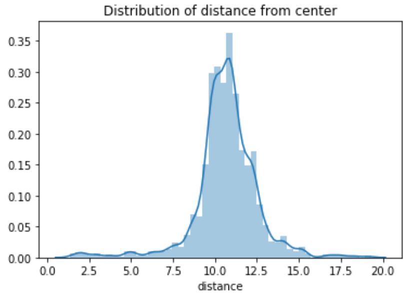
In the project a gave a look at the table but including the caracteristic is a quite heavy operation in term of data cleaning and therefore in this project it will not be included.

Neighborhood centers position and names are added to each neighbohoord via the Geopy interface.

Merging section

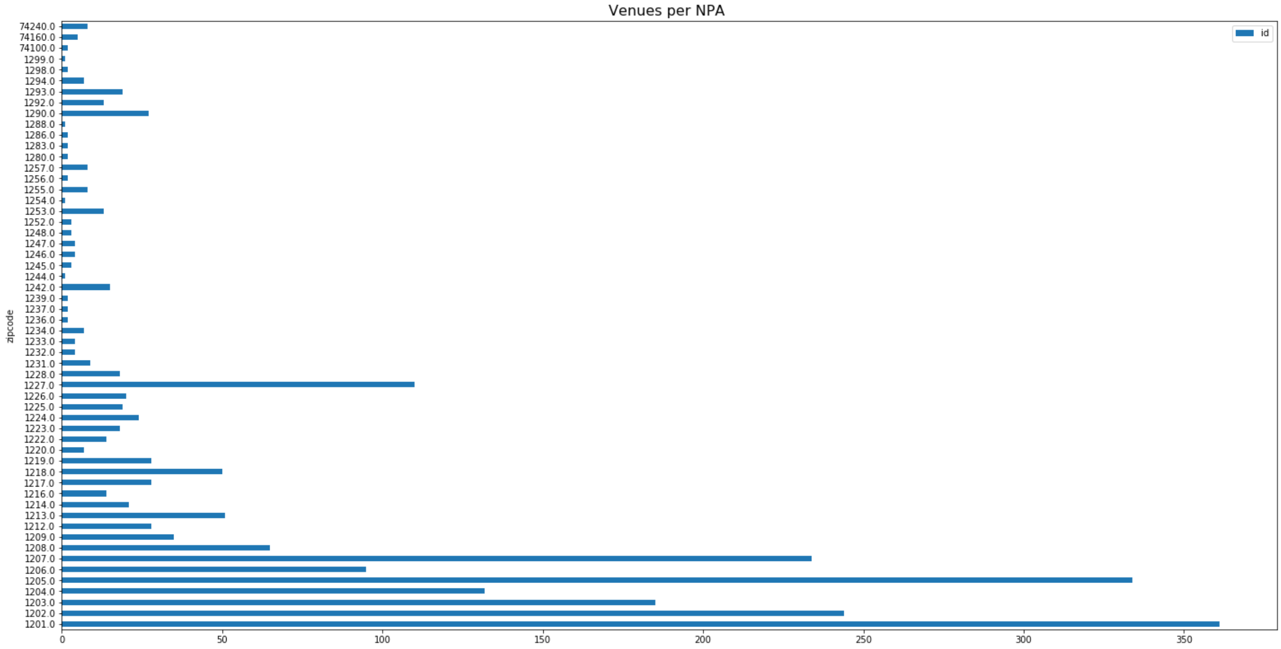
The final dataset comes from the merging of the cleaned listings and the neighborhood. It has to be noticed that some extensive work of cleaning is needed on different columns like the price to get them to be completely numerical and avoit text interference (e.g eliminate text caracters like $)

At this stage a predictor is added, the distance from center as that might be a factor influencing the amount of money people are getting on AirBnb

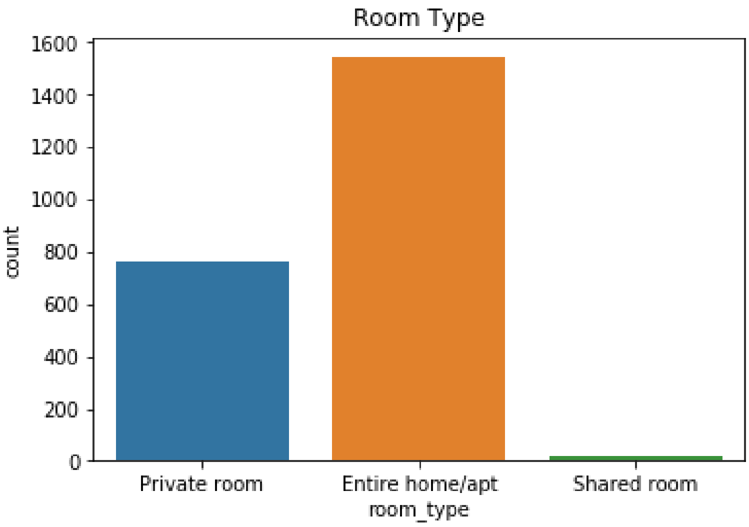


Exploratory Data Analysis

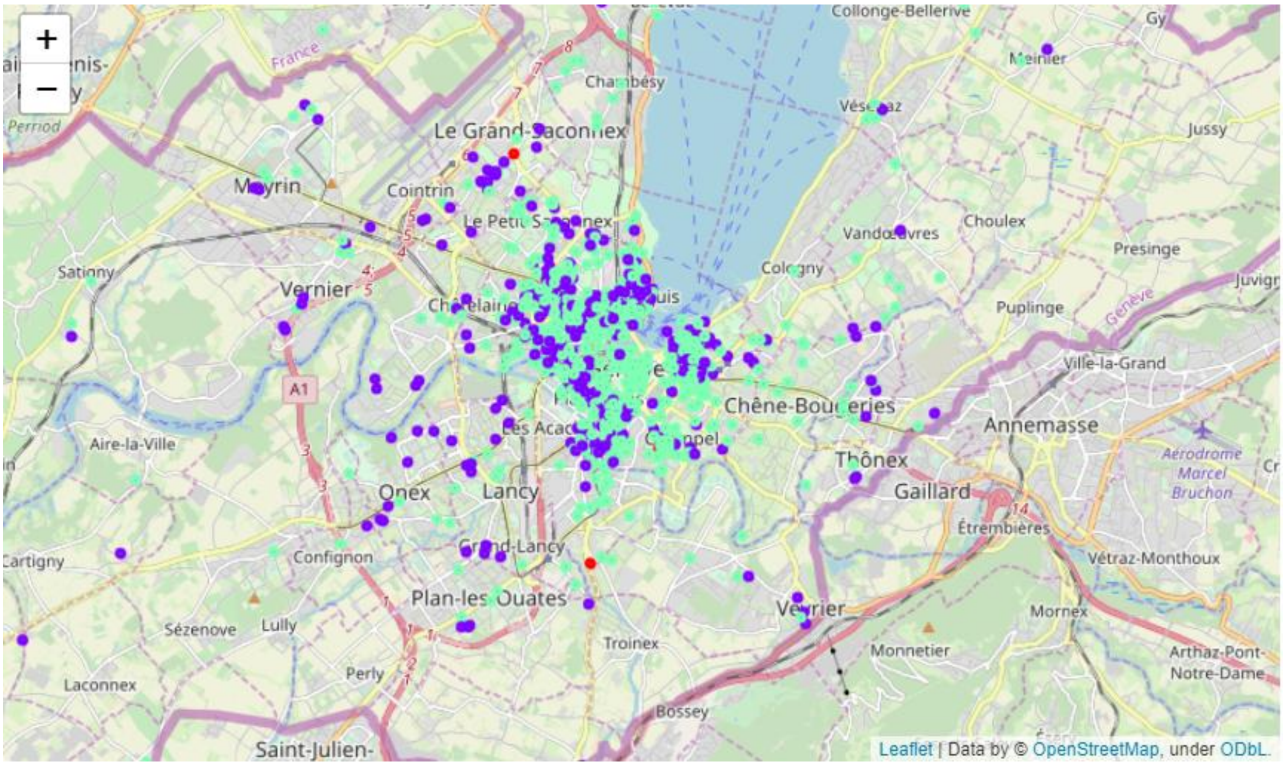
The distribution of venues per NPA can be reported here as well as their geographical distribution in Geneva



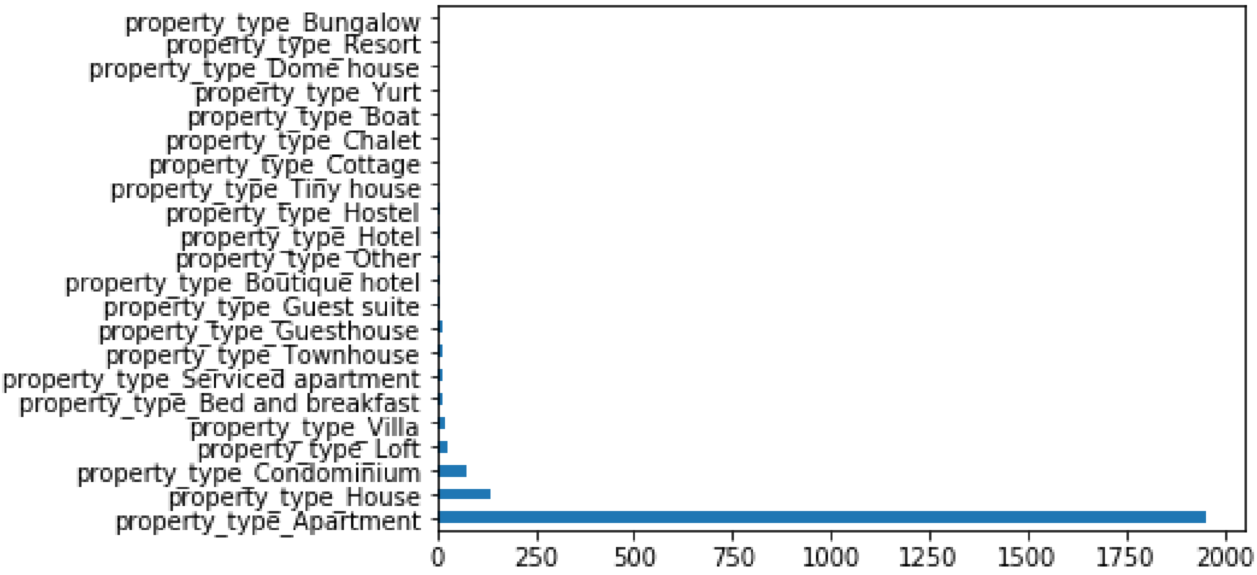
The room type is reported in the following graph



It is pretty interesting to notice the the most represented venues are entire homes or appartment and not single rooms. No shared rooms are present.



Most venues are in the city center.



The most commonly rented propriety is appartments and not houses.

Modeling :

The quantity to be forecasted has to be derived to what the site Inside Airbnb suggests

The following formula is applied :

$$Rev = 70\% \frac{N_{rewievs}}{50\%} \; Price \;\; Reviews_{montly} * 12 * MinLength$$

This approach has as hypothesis that 1 client every 2 leaves a review and the maximum availability is limited at 70%.

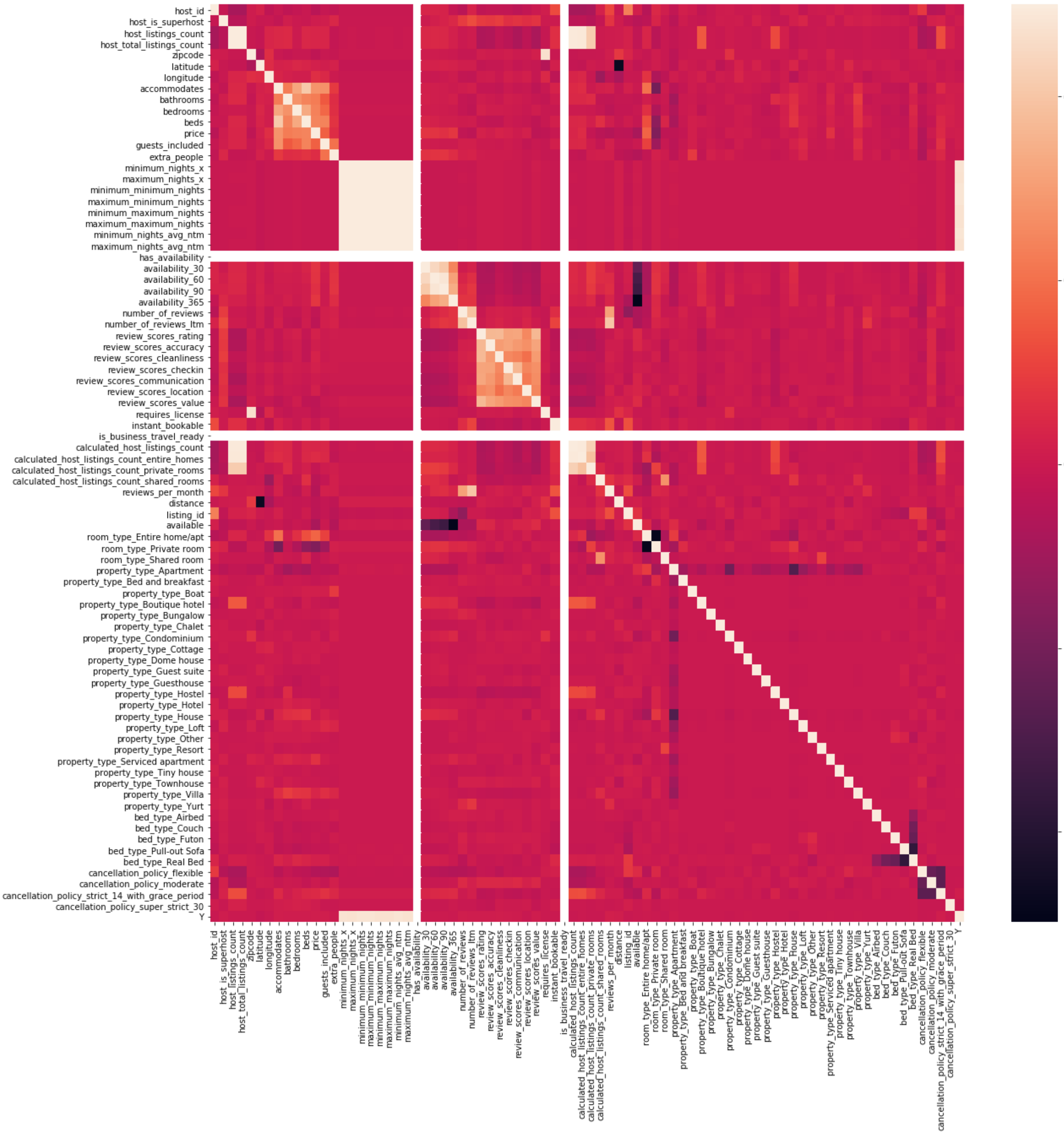
The distribution of revenues can be seen in the following graph

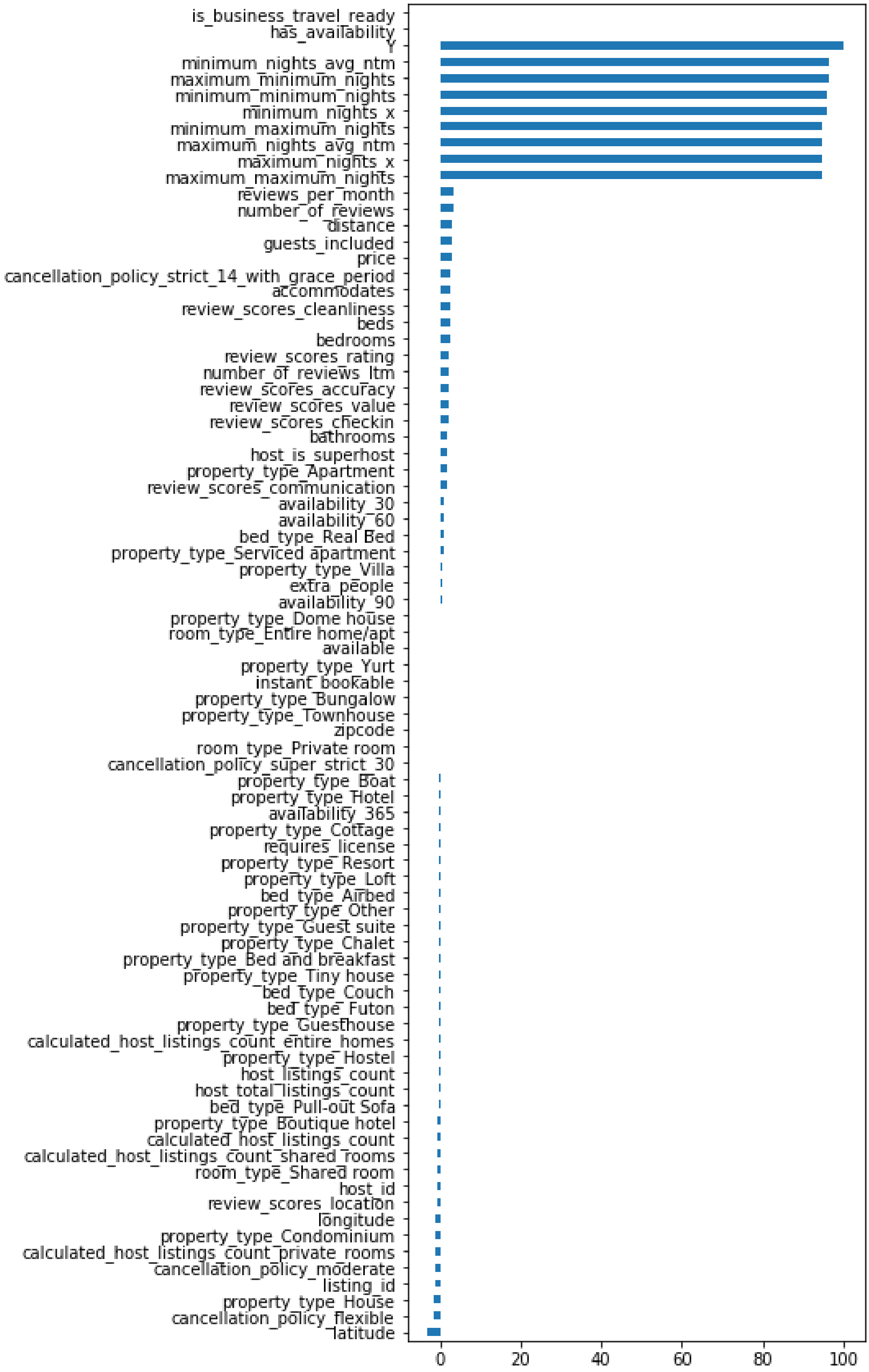
For modelling reason we take the logaritm of such quantity to get to something which is more handy for models.

The confusion matrix of the problem has this shape after habving computed the one hot encoding on some fetures to make them usable from models.

The most important insight is that some varibles are greatly correlated, expecially looking at the duration parameters which seem correlated with out targed variable.

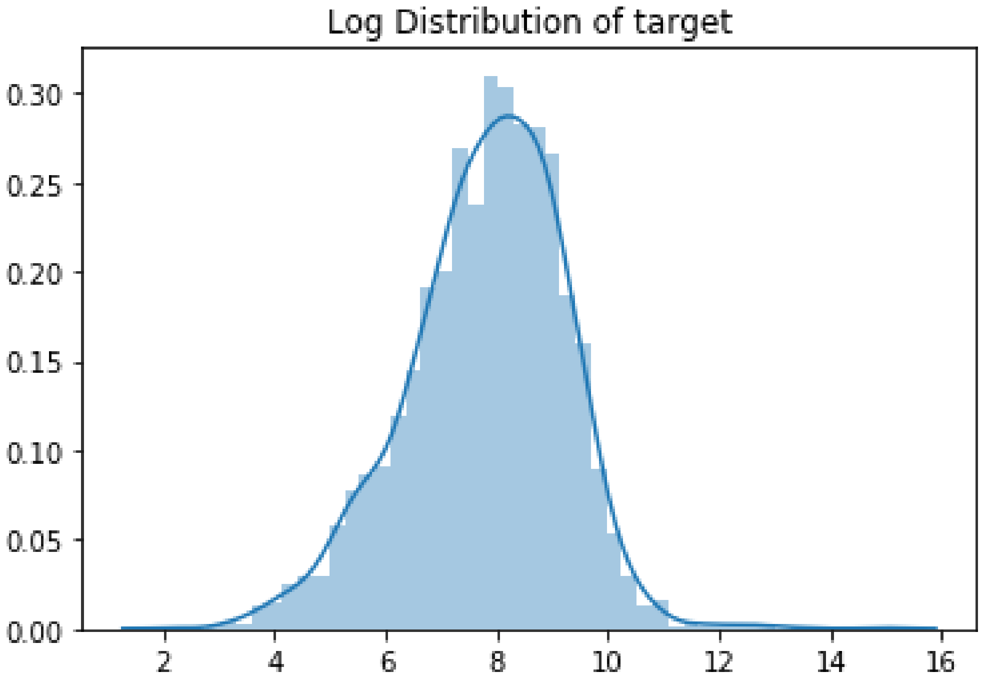
An extract of the correlation plots for Y is plotted here , it is pretty clear that parameters on the duration of the minimum night stay and the number of maximum lenght of the stay are the most important features.


host_id
host_is_superhost
host_listings_count
host_total_listings_count
zipcode
latitude
longitude
accommodates
bathrooms
bedrooms
beds
price
guests_included
extra_people
minimum_nights_x
maximum_nights_x
minimum_minimum_nights
maximum_minimum_nights
minimum_maximum_nights
maximum_maximum_nights
minimum_nights_avg_ntm
maximum_nights_avg_ntm
has_availability
availability_30
availability_60
availability_90
availability_365
number_of_reviews
number_of_reviews_ltm
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
requires_license
instant_bookable
is_business_travel_ready
calculated_host_listings_count
calculated_host_listings_count_entire_homes
calculated_host_listings_count_private_rooms
calculated_host_listings_count_shared_rooms
reviews_per_month
distance
listing_id
available
room_type_Entire home/apt
room_type_Private room
room_type_Shared room
property_type_Apartment
property_type_Bed and breakfast
property_type_Boat
property_type_Boutique hotel
property_type_Bungalow
property_type_Chalet
property_type_Condominium
property_type_Cottage
property_type_Dome house
property_type_Guest suite
property_type_Guesthouse
property_type_Hostel
property_type_Hotel
property_type_House
property_type_Loft
property_type_Other
property_type_Resort
property_type_Serviced apartment
property_type_Tiny house
property_type_Townhouse
property_type_Villa
property_type_Yurt
bed_type_Airbed
bed_type_Couch
bed_type_Futon
bed_type_Pull-out Sofa
bed_type_Real Bed
cancellation_policy_flexible
cancellation_policy_moderate
cancellation_policy_strict_14_with_grace_period
cancellation_policy_super_strict_30
Y
host_id
host_is_superhost
host_listings_count
host_total_listings_count
zipcode
latitude
longitude
accommodates
bathrooms
bedrooms
beds
price
guests_included
extra_people
minimum_nights_x
maximum_nights_x
minimum_minimum_nights
maximum_minimum_nights
minimum_maximum_nights
maximum_maximum_nights
minimum_nights_avg_ntm
maximum_nights_avg_ntm
has_availability
availability_30
availability_60
availability_90
availability_365
number_of_reviews
number_of_reviews_ltm
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
requires_license
instant_bookable
is_business_travel_ready
calculated_host_listings_count
calculated_host_listings_count_entire_homes
calculated_host_listings_count_private_rooms
calculated_host_listings_count_shared_rooms
reviews_per_month
distance
listing_id
available
room_type_Entire home/apt
room_type_Private room
room_type_Shared room
property_type_Apartment
property_type_Bed and breakfast
property_type_Boat
property_type_Boutique hotel
property_type_Bungalow
property_type_Chalet
property_type_Condominium
property_type_Cottage
property_type_Dome house
property_type_Guest suite
property_type_Guesthouse
property_type_Hostel
property_type_Hotel
property_type_House
property_type_Loft
property_type_Other
property_type_Resort
property_type_Serviced apartment
property_type_Tiny house
property_type_Townhouse
property_type_Villa
property_type_Yurt
bed_type_Airbed
bed_type_Couch
bed_type_Futon
bed_type_Pull-out Sofa
bed_type_Real Bed
cancellation_policy_flexible
cancellation_policy_moderate
cancellation_policy_strict_14_with_grace_period
cancellation_policy_super_strict_30
Y
0.8
0.4
0.0
−0.4
−0.8



The targe variable is transformed using a logaritms as its original distribution is quite skewed towards high values.

This is the visualisation of the problem target variable when already optimized.



Results

We use a RandomForestRegressor and the modeling pipeline is mande using a GridCV optimisation to tune the hyperparameters.

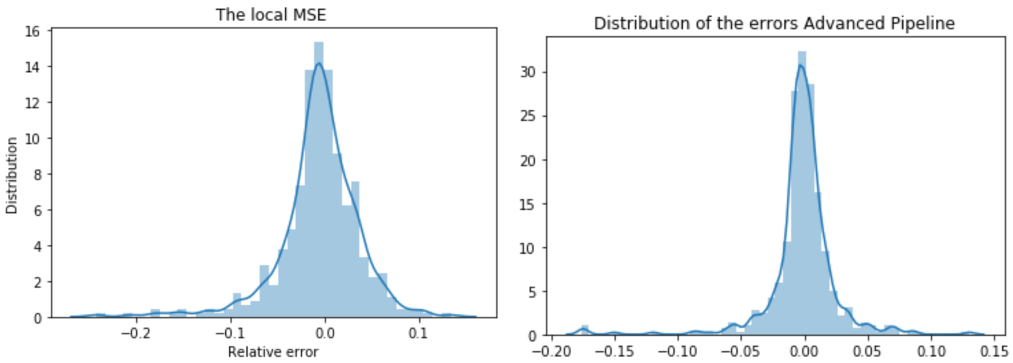
The optimal parameters are tuned and the model fitted we get to an overall non percentual non logaritmic error of : 21,75%

A similar procedure is performed using a pipeline and k-fold cross validaiton giving a slightly better result.

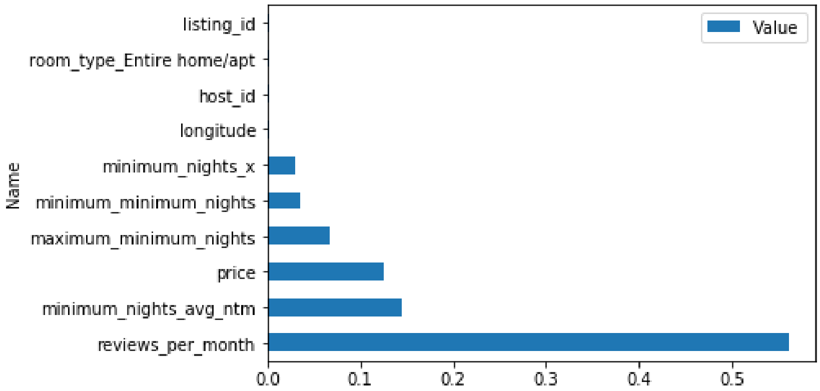
11.7%

The difference is in the hyperparameter tuning.

The distribution of the error is in the two cases :



The most important features are, in the second case :



As previously stated in the Correlatino matrix some factors actually predict almost all the variability. This is avery interesting fact.

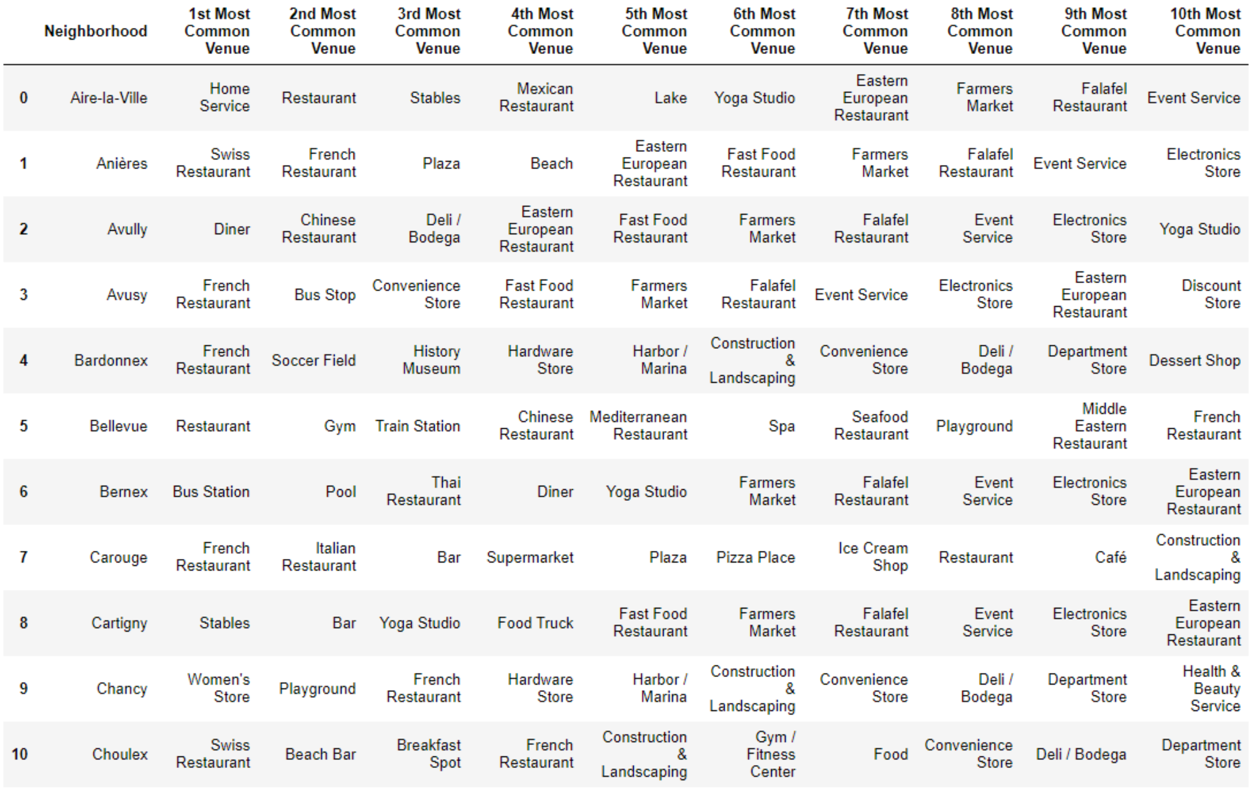
Improvement of the model

To improve the modelling caracteristics we imagine that venues in the zones of geneva might have an impact on the target variable

- Reseach venues on the Foursquare API

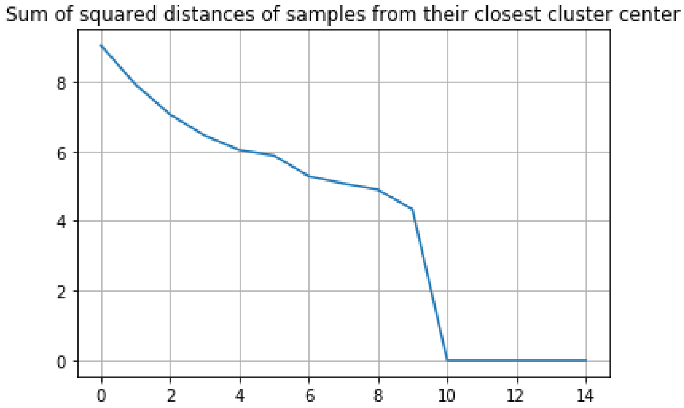
The procedure is similar to the one presented in the course.

I report the most common venues in different municipalities here:



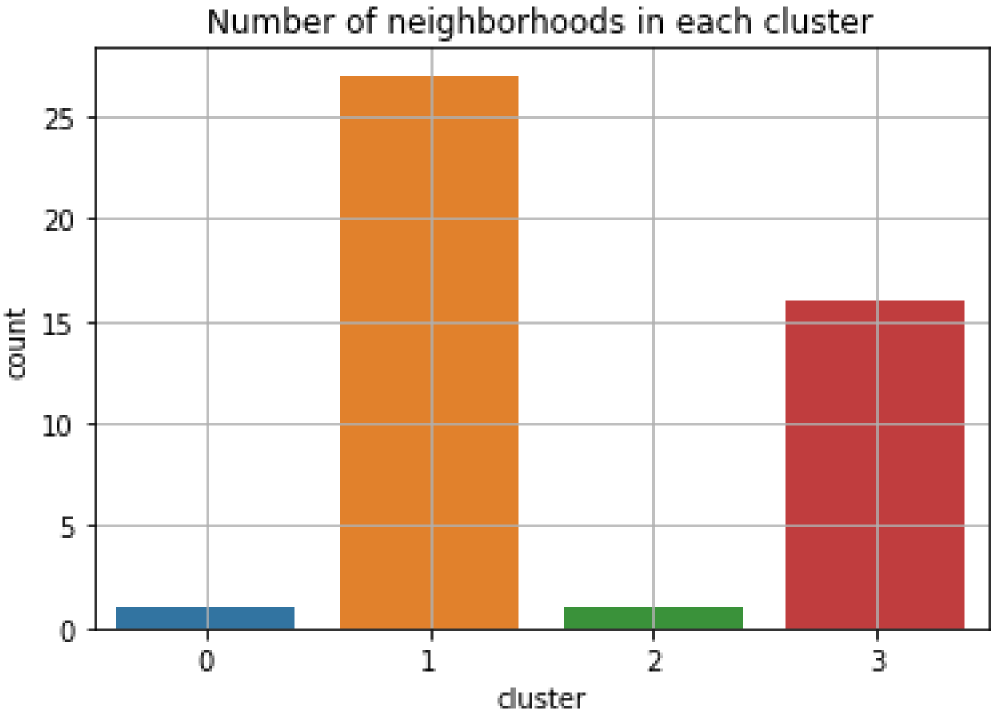
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aire-la-Ville	Home Service	Restaurant	Stables	Mexican Restaurant	Lake	Yoga Studio	Eastern European Restaurant	Farmers Market	Falafel Restaurant	Event Service
1	Anières	Swiss Restaurant	French Restaurant	Plaza	Beach	Eastern European Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Service	Electronics Store
2	Avully	Diner	Chinese Restaurant	Deli / Bodega	Eastern European Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Service	Electronics Store	Yoga Studio
3	Avusy	French Restaurant	Bus Stop	Convenience Store	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Service	Electronics Store	Eastern European Restaurant	Discount Store
4	Bardonnex	French Restaurant	Soccer Field	History Museum	Hardware Store	Harbor / Marina	Construction & Landscaping	Convenience Store	Deli / Bodega	Department Store	Dessert Shop
5	Bellevue	Restaurant	Gym	Train Station	Chinese Restaurant	Mediterranean Restaurant	Spa	Seafood Restaurant	Playground	Middle Eastern Restaurant	French Restaurant
6	Bernex	Bus Station	Pool	Thai Restaurant	Diner	Yoga Studio	Farmers Market	Falafel Restaurant	Event Service	Electronics Store	Eastern European Restaurant
7	Carouge	French Restaurant	Italian Restaurant	Bar	Supermarket	Plaza	Pizza Place	Ice Cream Shop	Restaurant	Café	Construction & Landscaping
8	Cartigny	Stables	Bar	Yoga Studio	Food Truck	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Service	Electronics Store	Eastern European Restaurant
9	Chancy	Women's Store	Playground	French Restaurant	Hardware Store	Harbor / Marina	Construction & Landscaping	Convenience Store	Deli / Bodega	Department Store	Health & Beauty Service
10	Choulex	Swiss Restaurant	Beach Bar	Breakfast Spot	French Restaurant	Construction & Landscaping	Gym / Fitness Center	Food	Convenience Store	Deli / Bodega	Department Store

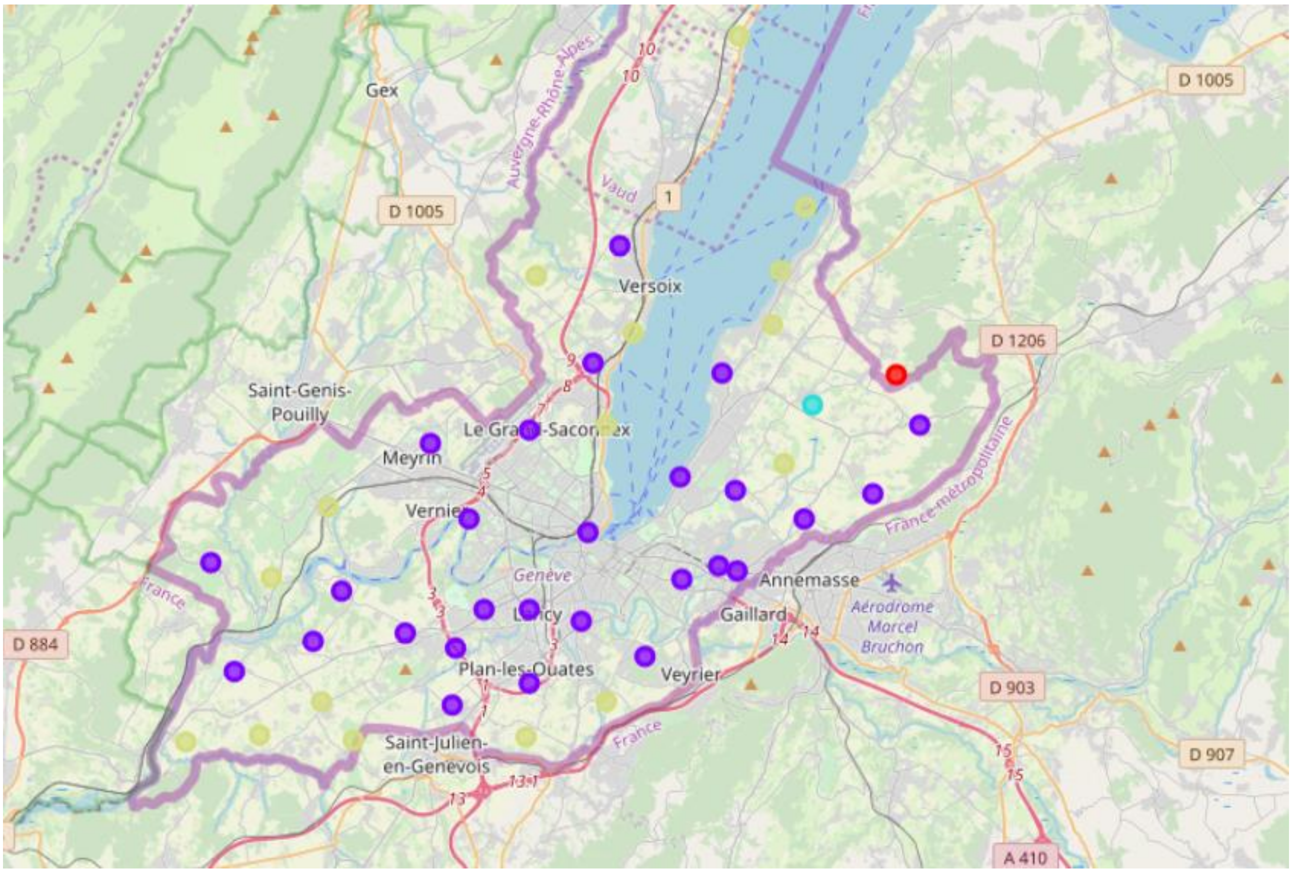
- The clustering results bring us to choose a number of clusters of 4 as it seems a reasonable tradeoff between the difference introduced by the clusters and the complexity which is needed to add.
-



-
- Recompute the pipeline for modelling.

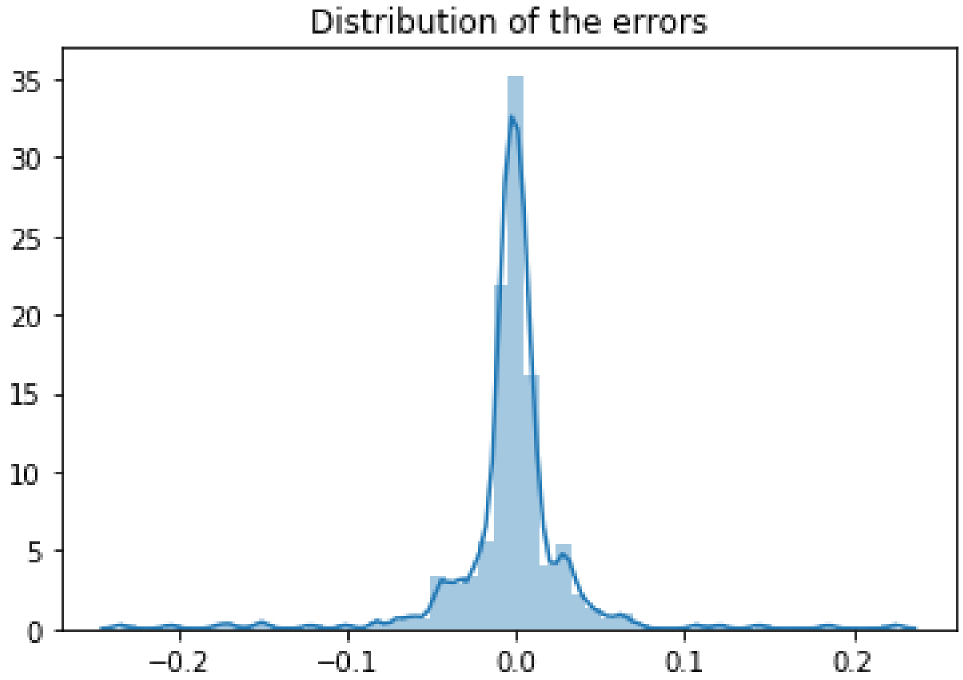
Interesting is the distribution of the different clusters made by the neighborhoods which gives us some more peripherical villages in cluster number 3 and the more central in cluster number 1





The most important conclusion of this process is that the results of the modelling do not change significntly using the new scenario and the feature importances are not changed.

Clusters are added as dimensions to the random forest regressor so as to improve the quality of the modelling.



The score on the non logaritmic test sample is 12.70% which is higher than the previous with the best pipeline.

The model does not like to have more variable to choose from and the benefit of informatio from clusrer is lower than the noise introduced.

```
12.70656646677726 %
```

Supporting this thesis there is also the fact that the most 15 important features do not change.

Conclusion

A model to forecast the revenues from AirBnb is done in this project.

The results show a results of XXX percentage on the annual revenues.

The visual statistics can provide useful insights for decision makers of a city to undesrant if the overall amount of money is disturbing the normal renting activity in a city as well as potential renters to have an idea of such market

- in geneva more than 2000 apartments are rented on airbnb every year.

- The most influencial parameters for detemining the revenues are the reviews, as stated in the AIRBNB the number of reviews tells if a renter location is good and y the owned might get ot of money out of it.

- Adding the cluster information to the model, based on the caracteristics onthe neighborhood, does not increase accuracy but it risks to lower it adding noise in the model. The reason might be that taking information for quite a big area might not be representative for thesingle points.

- another reason might be that geneva is a geographically small canton and differences are quite small among different municipalities in the canton.

Further exploration might be needed to :

- compare results from geneva with other cities in the world or in Switzerland
- forecast a price in absence of reviews
- get to some information about the photos of the houses
- include information from the text from reviews.

This work is intended to be open source and is based on publicly available data.

Patrizio Canzi