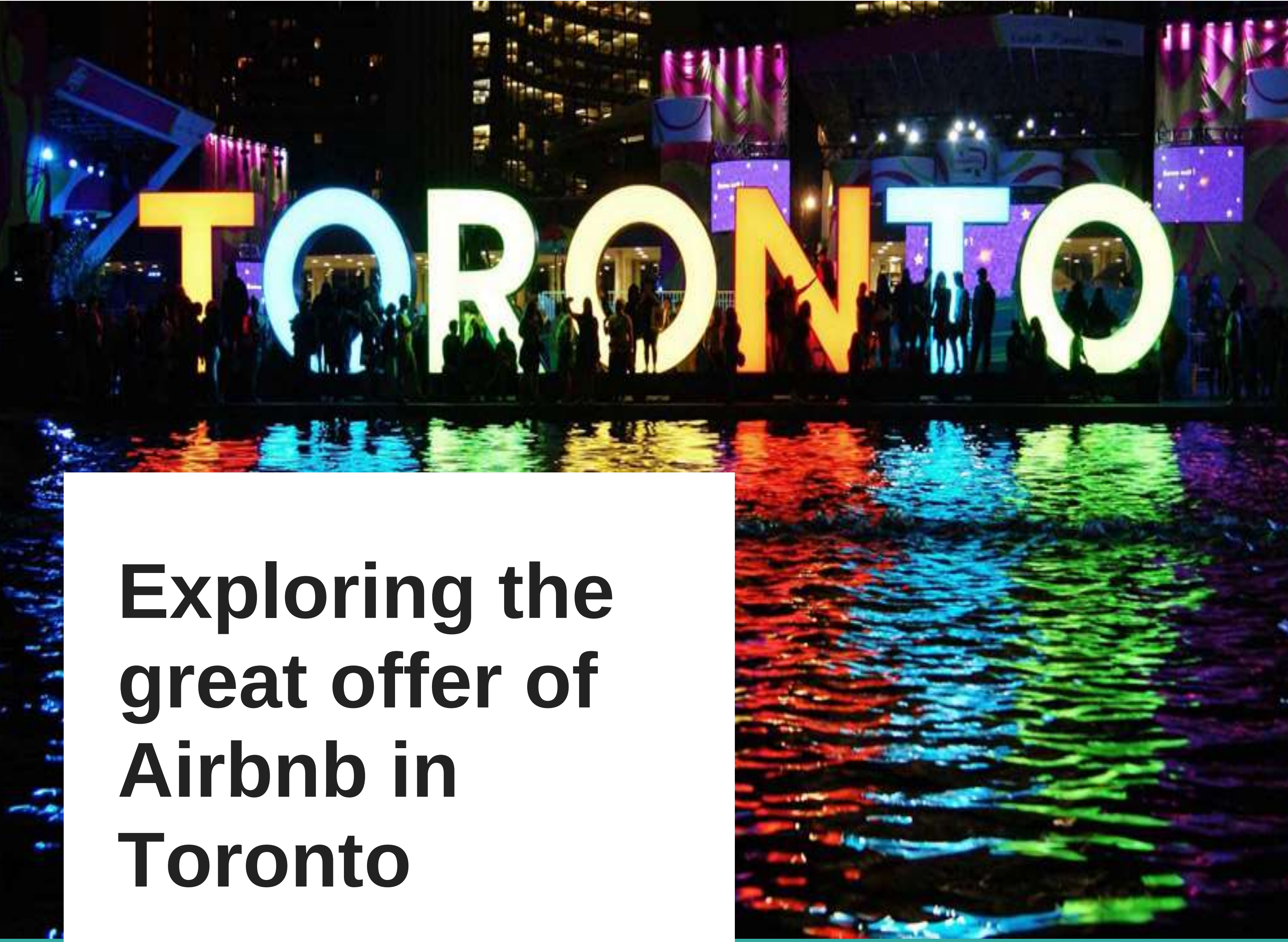


Exploring the great offer of Airbnb in Toronto

September 24, 2019

Applied Data Science Capstone
Created by: Israel Sumano

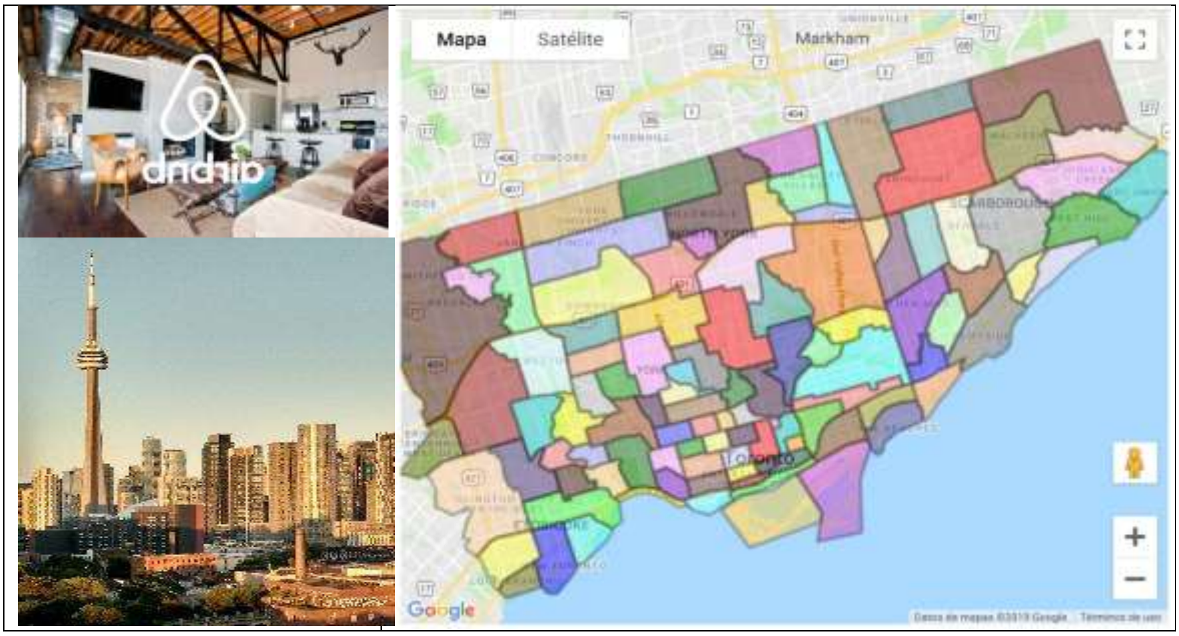
Exploring the great offer of Airbnb in Toronto

Problem Background

As part of the final IBM Capstone project, we will carry out an analysis that a data scientist could perform in his professional practice when faced with a real problem. The objectives of this project will be to define a problem, search for data in the available media (commonly on the internet and use the resources provided by Foursquare to enrich the analysis of the different neighborhoods of the selected city. For this case, I will choose the city of Toronto Ontario, Canada.

Toronto is a city considered multicultural, with an excellent standard of living, tourist attractions, universities and institutions for learning languages and technology. This tourist and academic offer generates an important demand of places for lodging of all kinds, either for students with stays of 90 days up to 6 months, as for sporadic tourists who only seek to stay a couple of nights. Currently, the Airbnb platform provides additional accommodation options to the normal offer of traditional hotels, guest houses and hostels.

This means that the market is highly competitive. As it is a highly quoted destination in Canada, the cost of staying in Toronto (lodging, food, transportation, amenities, etc.) is considerable. Therefore, it is vital to have good information to choose a place of accommodation that allows to balance the expenses of stay. Additionally, the thousands of offers available on the Airbnb platform make the search a stressful challenge, and sometimes, subject to "good luck".



Problem Description

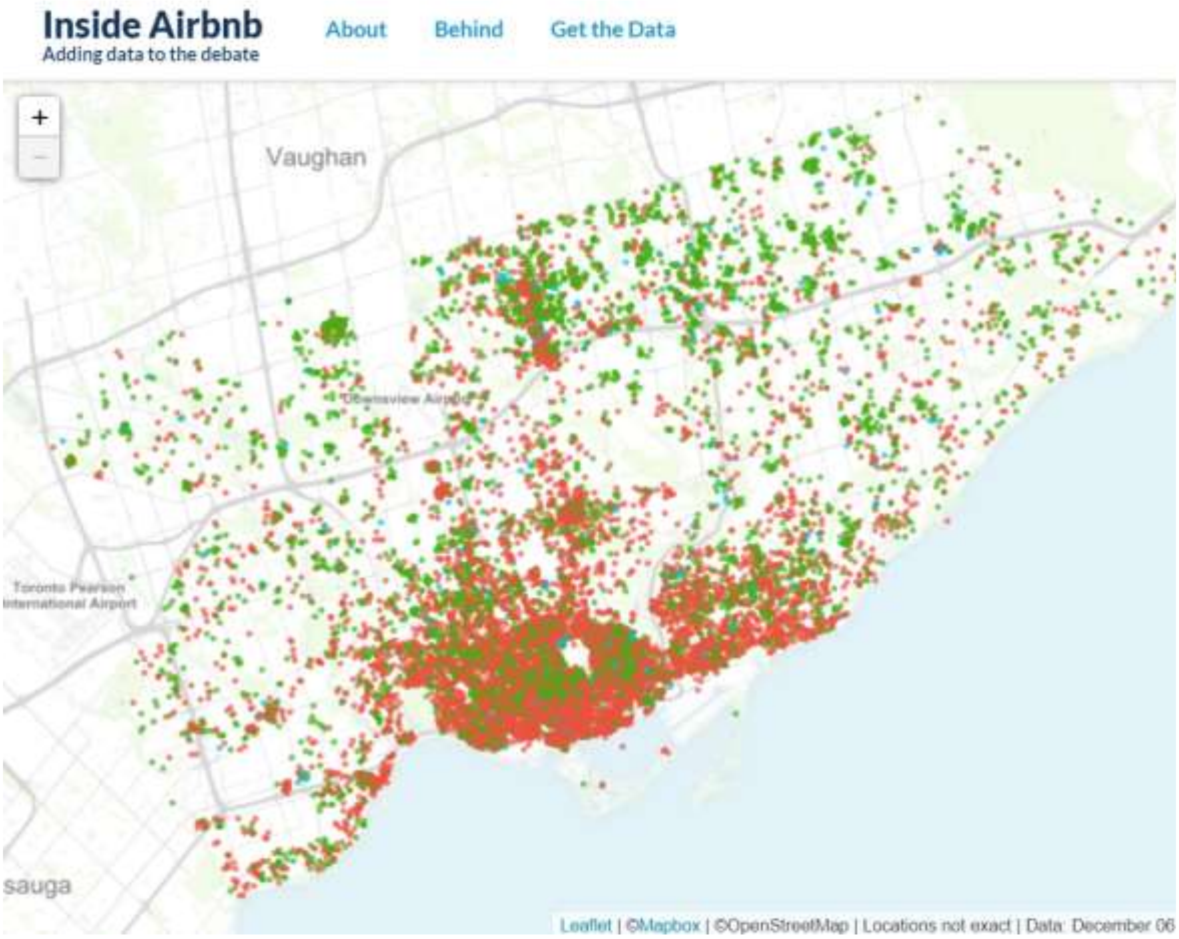
The basis of this project aims to help the community of Airbnb users, mainly exchange students, tourists interested in Toronto as a vacation destination, or even a potential entrepreneur who wants to become a “host” to have a useful tool that allows them to process the considerable volume of data to "discover":

- What are the different types of properties available in this city?
- What are neighborhoods have the most expensive listing prices on average?
- What are the average prices by type of property for accommodation?

And even help an entrepreneur interested in promoting accommodation in Airbnb to know:

- What would be a competitive price to offer? According to your type of property and neighborhood where it is located.

This project also seeks to take advantage and extend the experience already acquired in previous years and enrich it with data from the Airbnb bases.


Inside Airbnb
Adding data to the debate
About
Behind
Get the Data
+
−
Vaughan
Downsview Airport
Toronto Pearson International Airport
sauga
Leaflet | ©Mapbox | ©OpenStreetMap | Locations not exact | Data: December 06

Data acquisition

The basis of this project aims to help the community of Airbnb users, mainly exchange students, tourists

Data 1 :

The source of data for building this report was:

http://data.insideairbnb.com/canada/on/toronto/2019-08-08/data/listings.csv.gz

This csv file is date compiled 08 August, 2019. **It contains 21617 rows and 106 columns**

This data set was the main source for analyzing and building this report.

Data 2 :

The source of our data is in the following link

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

so it will be necessary to extract its content, supported by BeautifulSoup libraries.

There are:

• 180 Postal codes • 11 Boroughs • 208 Neighborhoods

Obviously, it will be necessary to clean and choose certain neighborhoods, as we integrate the data obtained from Airbnb

Data 3 :

We will use the Foursquare API to explore neighborhoods in the city of Toronto. Additionally, the Foursquare API data will help us to show the locations of the accommodations and obtain additional information about the neighborhoods (restaurants, hotels, amenities, etc.)

Metodology

In data science, an excellent way to explore a dataset is try to visualize some trends on the main characteristics according to the problem context that we want to solve, in this case the range of accommodation price offer and identify a competitive price**.**
However, it is necessary to go through a series of stages to have the data sufficiently clean to carry out a reliable analysis. The stages considered for this study were:

1. Data Manipulation
2. Feature Selection
3. Missing Values Handling
4. Data visualization
5. Outlier Removal
6. Data Blending
7. Normalization
8. Partitioning
9. Data Modeling
10. Evaluating results

Exploring original data source

The original data source from Airbnb contains 21617 rows and 106 columns. Therefore, it was necessary to select only the characteristics were considered important for the current analysis (and even future analyzes). The features chosen are:

1. 'id'
2. 'listing_url'
3. 'street'
4. 'neighbourhood'
5. 'neighbourhood_cleansed'
6. 'neighbourhood_group_cleansed'
7. 'city'
8. 'state'
9. 'zipcode'
10. 'country'
11. 'latitude'
12. 'longitude'
13. 'property_type'
14. 'room_type'
15. 'bedrooms'
16. 'beds'
17. 'bed_type'
18. 'price'
19. 'weekly_price'
20. 'monthly_price'
21. 'minimum_nights'
22. 'maximum_nights'
23. 'has_availability'
24. 'availability_365'
25. 'number_of_reviews'
26. 'last_review'
27. 'review_scores_rating'
28. 'review_scores_cleanliness'
29. 'review_scores_location'
30. 'reviews_per_month'

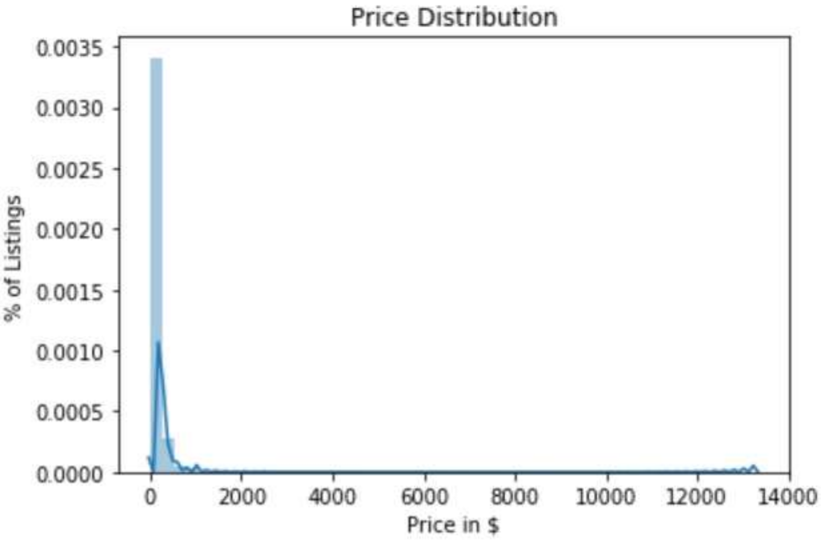
Checking how many "Null" or "NaN" values are in all our data

Missing more than 15% {'reviews_per_month', 'last_review', 'review_scores_rating', 'review_scores_location', 'review_scores_cleanliness', 'neighbourhood_group_cleansed', 'monthly_price', 'weekly_price'}
Missing more than 25% {'monthly_price', 'weekly_price', 'neighbourhood_group_cleansed'}
Missing more than 50% {'monthly_price', 'weekly_price', 'neighbourhood_group_cleansed'}

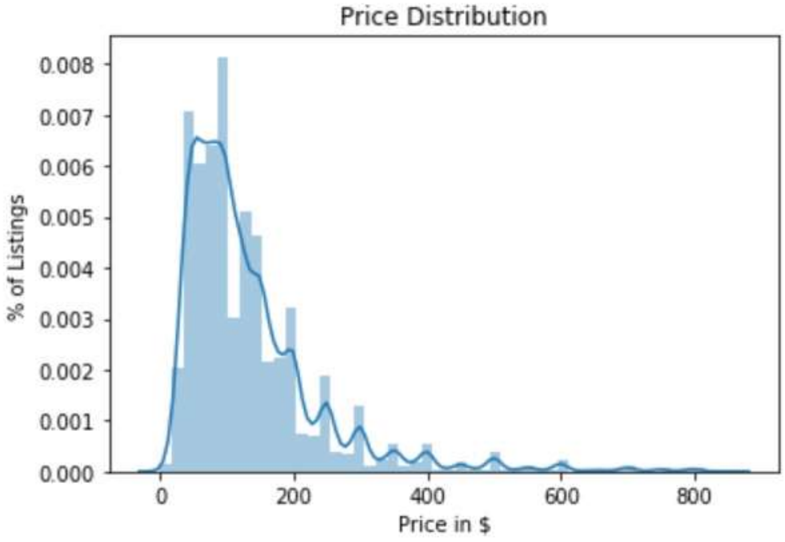
Fortunately, the most relevant columns for this analysis process were not detected in this null value inspection.

Data visualization, with the current set

First, with the distribution graph we identify the level of outliers in the data.

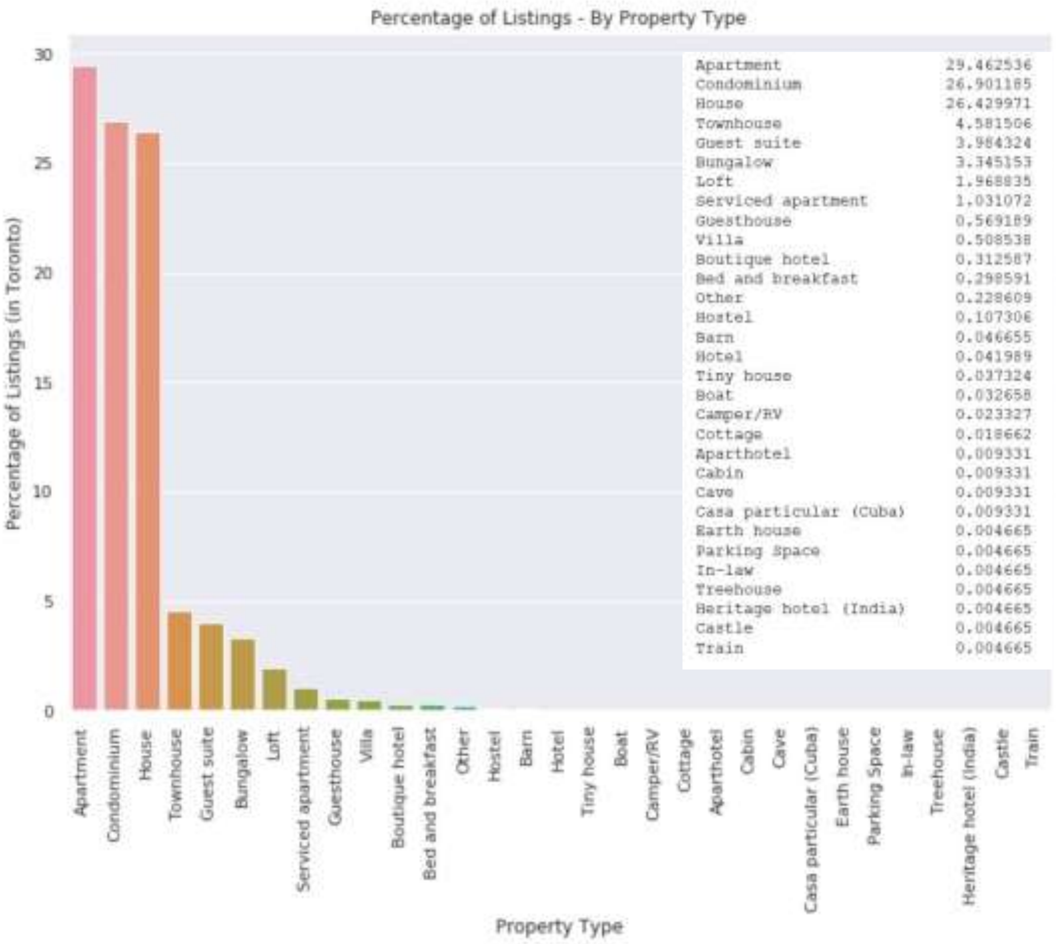


After remove *outliers* (listings priced more than 3x standard deviation), this step is very important. The distribution graph without *outliers* is:



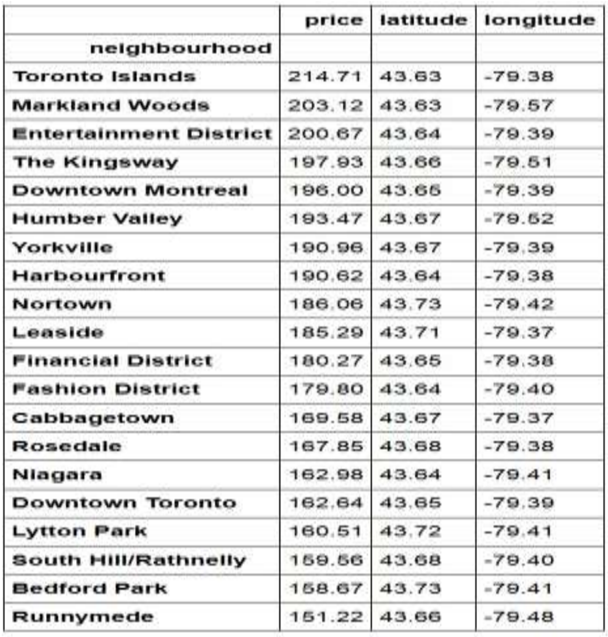
In accordance with the data set without outliers, the analysis is carried out to begin answering the questions that motivate this study.

What are the different types of properties available in Toronto

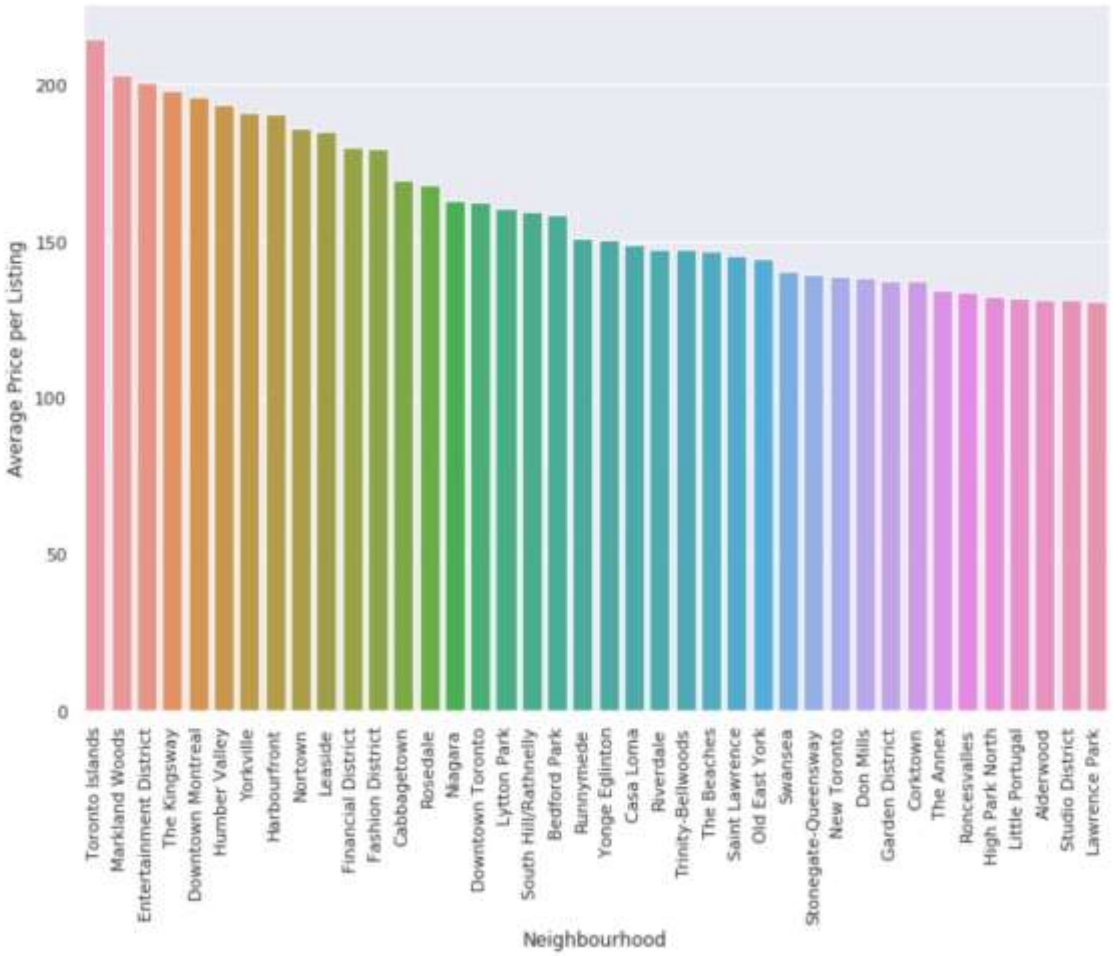


```
Apartment                 29.462536
Condominium               26.901185
House                     26.429971
Townhouse                  4.581506
Guest suite                3.984324
Bungalow                   3.345153
Loft                       1.968835
Serviced apartment         1.031072
Guesthouse                 0.569189
Villa                      0.508538
Boutique hotel             0.312587
Bed and breakfast          0.298591
Other                      0.228609
Hostel                     0.107306
Barn                       0.046655
Hotel                      0.041989
Tiny house                 0.037324
Boat                       0.032658
Camper/RV                  0.023327
Cottage                    0.018662
Aparthotel                 0.009331
Cabin                      0.009331
Cave                       0.009331
Casa particular (Cuba)     0.009331
Earth house                0.004665
Parking Space              0.004665
In-law                     0.004665
Treehouse                  0.004665
Heritage hotel (India)     0.004665
Castle                     0.004665
Train                      0.004665
```

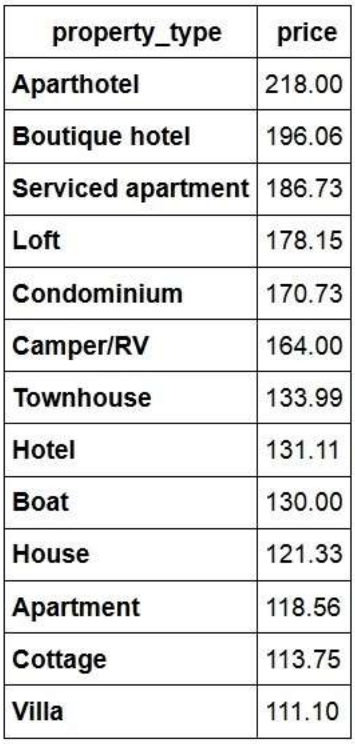
What are neighborhoods have the most expensive listing prices on average?



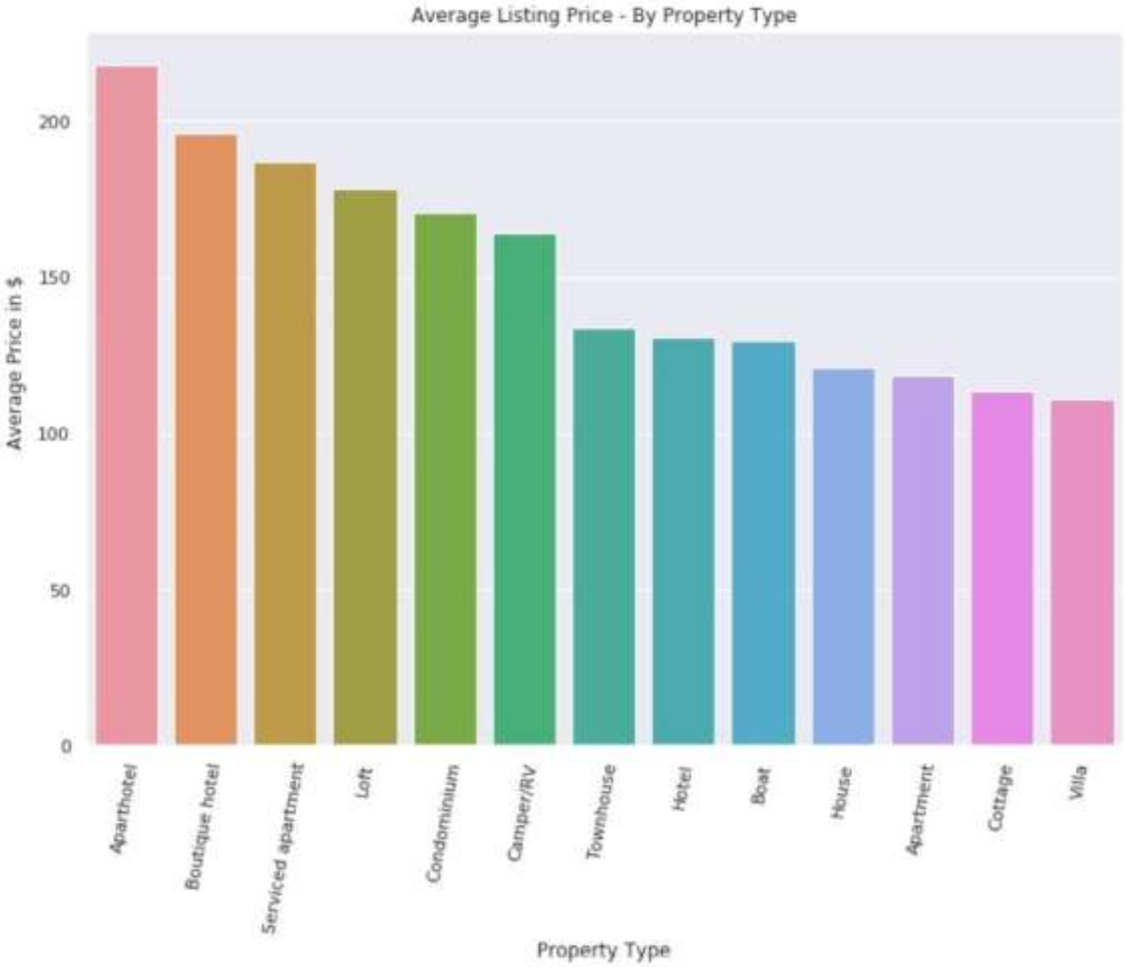
	price	latitude	longitude
neighbourhood			
Toronto Islands	214.71	43.63	-79.38
Markland Woods	203.12	43.63	-79.57
Entertainment District	200.67	43.64	-79.39
The Kingsway	197.93	43.66	-79.51
Downtown Montreal	196.00	43.65	-79.39
Humber Valley	193.47	43.67	-79.52
Yorkville	190.96	43.67	-79.39
Harbourfront	190.62	43.64	-79.38
Nortown	186.06	43.73	-79.42
Leaside	185.29	43.71	-79.37
Financial District	180.27	43.65	-79.38
Fashion District	179.80	43.64	-79.40
Cabbagetown	169.58	43.67	-79.37
Rosedale	167.85	43.68	-79.38
Niagara	162.98	43.64	-79.41
Downtown Toronto	162.64	43.65	-79.39
Lytton Park	160.51	43.72	-79.41
South Hill/Rathnelly	159.56	43.68	-79.40
Bedford Park	158.67	43.73	-79.41
Runnymede	151.22	43.66	-79.48



Which property types are the most expensive in Toronto?



property_type	price
Aparthotel	218.00
Boutique hotel	196.06
Serviced apartment	186.73
Loft	178.15
Condominium	170.73
Camper/RV	164.00
Townhouse	133.99
Hotel	131.11
Boat	130.00
House	121.33
Apartment	118.56
Cottage	113.75
Villa	111.10

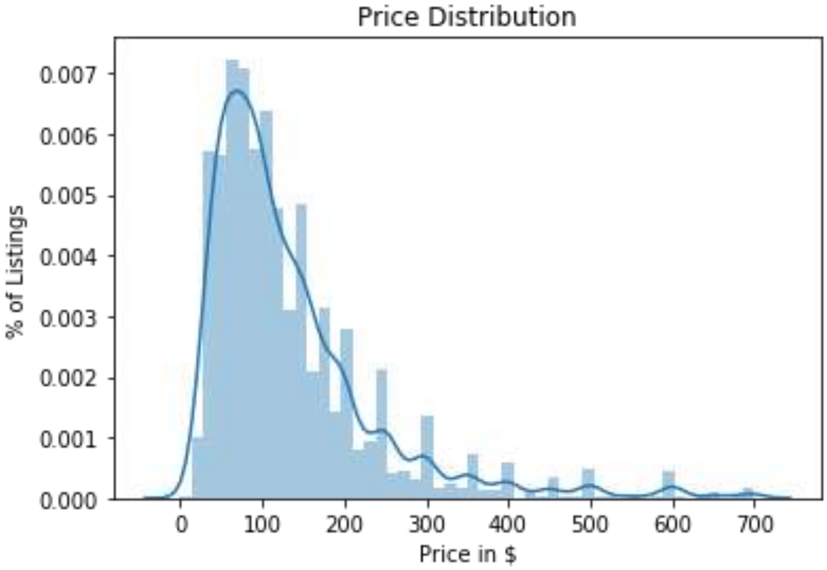


What would be a competitive price to offer?

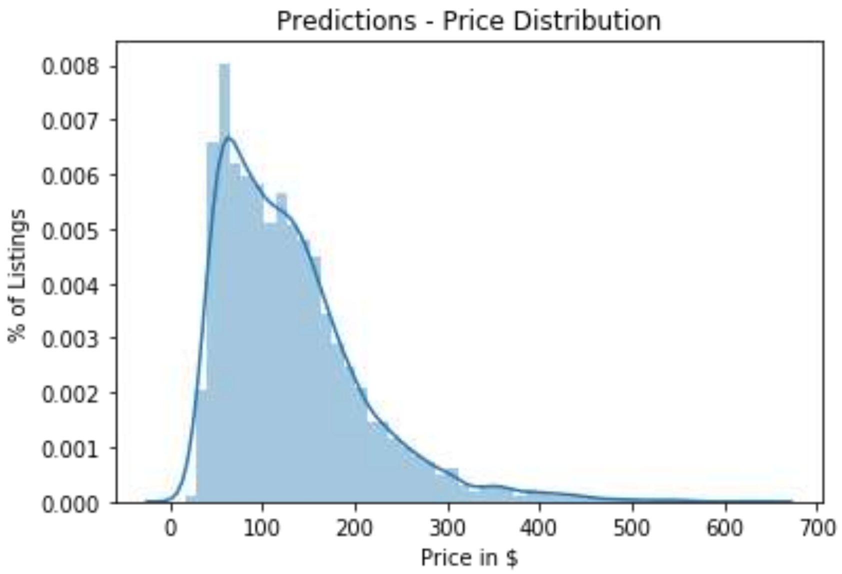
Data Modeling with linear regression

To obtain information about the factors that influence the price, we construct a linear regression model to then inspect the coefficients of that model.

This is the distribution plot for the true prices in our test set. See how the fitted line is not perfectly smooth.



This graph is the distribution plot of the predicted prices (with Linear Regression).
The model over estimates the number of properties in the $100 to $200 range. The model is also under estimating the number of higher price listings.



The R2 score is around 0.48 when using a Linear Regression model. In general, the higher the R-squared, the better the model fits your data.
In this case, R2=0.48 may not be the best score. But analyzing the graph we can consider that it is not bad.

Final Results

Next, the summary table with the final results (the most significant) of the study.

What are the different types of properties available in this city?	
Apartment	29.462536
Condominium	26.901185
House	26.429971
Townhouse	4.581506
Guest suite	3.984324
Bungalow	3.345153
Loft	1.968835
Serviced apartment	1.031072

What are neighborhoods have the most expensive listing prices on average?	
Toronto Islands	$214.71
Markland Woods	$203.12
Entertainment District	$200.67
The Kingsway	$197.93
Downtown Montreal	$196.00
Humber Valley	$193.47
Yorkville	$190.96
Harbourfront	$190.62
Nortown	$186.06
Leaside	$185.29
Financial District	$180.27
Fashion District	$179.80

What are the average prices by type of property for accommodation?

Aparthotel	$218.00
Boutique hotel	$196.06
Serviced apartment	$186.73
Loft	$178.15
Condominium	$170.73
Camper/RV	$164.00
Townhouse	$133.99
Hotel	$131.11
Boat	$130.00
House	$121.33
Apartment	$118.56
Cottage	$113.75
Villa	$111.10

What would be a competitive price to offer? According to your type of property and neighborhood where it is located.

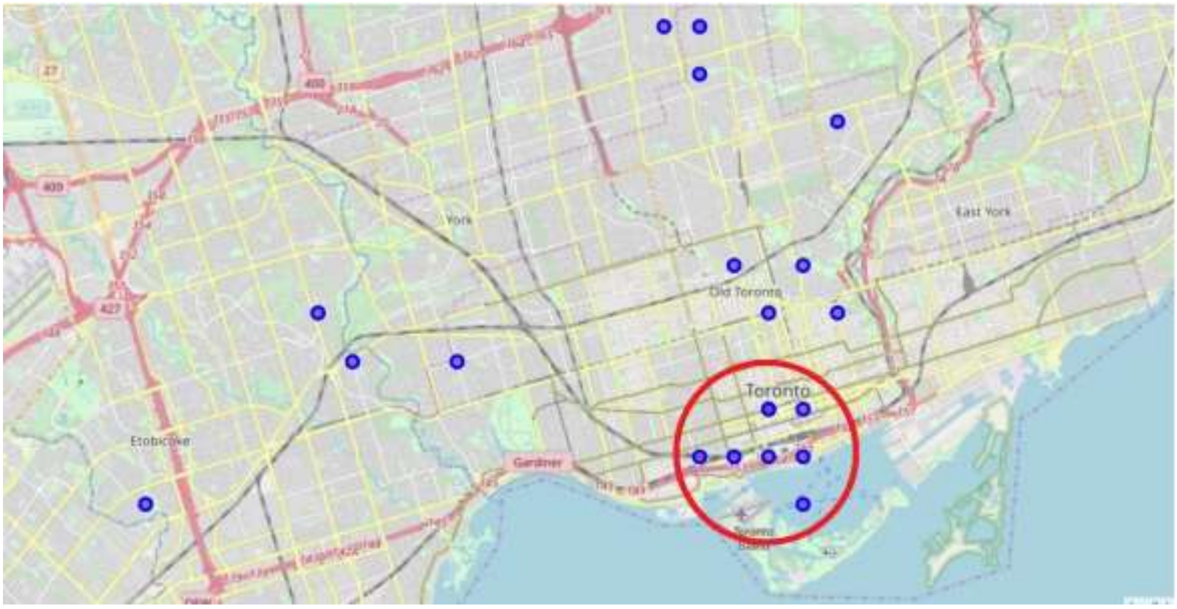
A competitive price, regardless of the type of property, is around $ 148.00
Considering this price, the type of property offered can be a Condominium o Townhouse

Discussion of the results and observations

Given the volume of accommodation offers, it is difficult to restrict yourself to a type of property and neighborhood.

A curious aspect in the study was to discover that the top 20 neighborhoods with the highest prices per night for accommodation are very scattered throughout the city.

But it seems that is attractive to be in the vicinity of Toronto Islands and entertainment district; as the following map shows:



Conclusion

In conclusion, in this new technological era, where data floods us day by day, the important thing is not only to know those data, but to analyze them to know how to use them in each case.

Our challenges as IT professionals and data scientists are:

- Be able to analyze and interpret the results of a specific problem and the ability to explain these results to provide an objective and rigorous vision of all the processes of a business.
- Determine how to obtain value from the data obtained and the clear definition of Data Science Strategy.
- Get the perfect alignment between cloud services, Cloud Computing, and Data Science.
- Transform decision making using state-of-the-art automation technologies.