Airbnb Toronto Price Prediction

Capstone Project Presentation

UofT SCS Data Analytics Boot Camp
Aug 10th, 2022

Contributors:
Carolina Semerano
Sukanya Ghosh
Yesha Tharwala

[Link to Project Presentation]

Contents

Overview

Airbnb is an online marketplace for short term rentals. Airbnb allows people from all over the world to host their homes as someone's next stay.

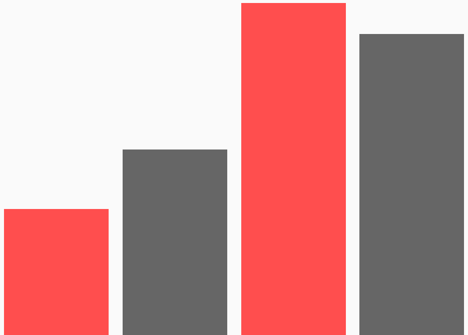
Properties can range from houses, apartments to single and shared rooms and are priced per night or per stay.

The problem

With the ever changing market it can be challenging for Airbnb hosts to determine the optimal rent prices for their properties.

Since price depends on numerous factors ranging from property type to amenities offered, as well as location customer reviews and ratings.

Hosts can often misinterpret the prices for their neighbourhoods and miss the opportunity to good profits for their listings.



The solution

Construct a data driven solution by using machine learning to predict rental prices for each property.

Data Source

The file clean_data_bourgh.csv, contains data about Airbnb listings in Toronto, Canada. The dataset contains a total of 15171 records, where each row represents a unique listing and every column represents important data about the listing. The following is the description of some columns in this dataset.

1. **host_since**: The date that host listed their first Airbnb listing.
2. **host_response_rate**: How fast the host responded to customer inquiries.
3. **neighbourhood**: The Toronto neighbourhood of the listing.
4. **property_type**: Type of property (Entire home, private room, share room, hotel room)
5. **price**: Price of the listing property per day
6. **bourgh**: The Toronto bourgh where the of the listing property

Data Cleaning

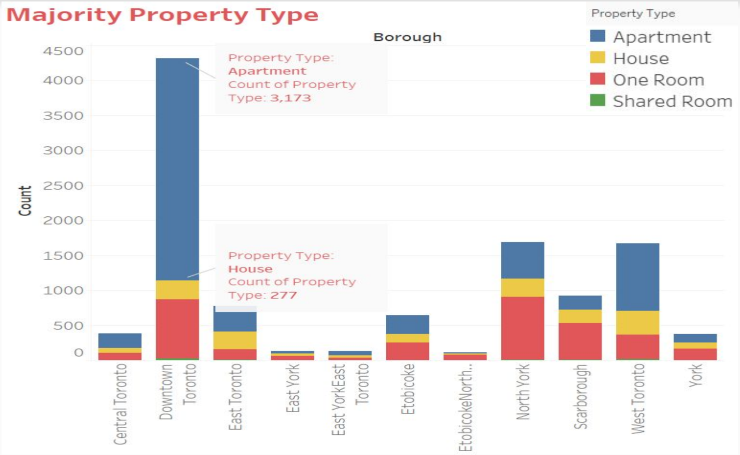
Cleaned the original listings.csv file using Pandas library and created clean_data_borough.csv for project analysis

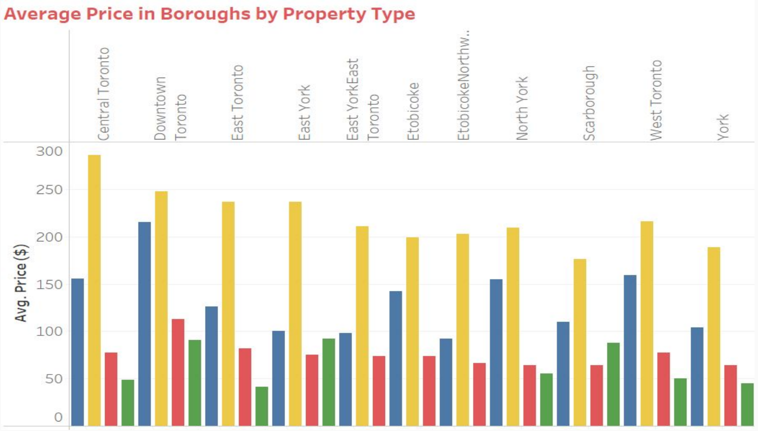
1. Dropped 37 columns that are not relevant for project analysis
2. Scraped Toronto Postal Codes from Wikipedia to determine the relation between postal codes, neighbourhood and boroughs
3. Added postal codes and bourgh data to original dataset to match each neighbourhood with a bourgh

Exploration Data Analysis

The data exploration phase of the project is conducted in Python and Tableau. We primarily analyzed the data based on four segments, host details, location, reviews and amenities. Within the host details segment our goal was to determine whether factors like being a superhost, having a verified identity and the number of listings the host has in the city has any impact on average prices. In the location segment we derived insights like the most and least expensive neighbourhoods in the city, the average prices in each Toronto borough, etc. The reviews segment we dived a little deeper into understanding the impact customer ratings and reviews have on prices and popularity. Lastly, we looked into popular amenities and whether or not they are offered by most Airbnb listings and the impact they have on average prices.

Exploration Data Analysis

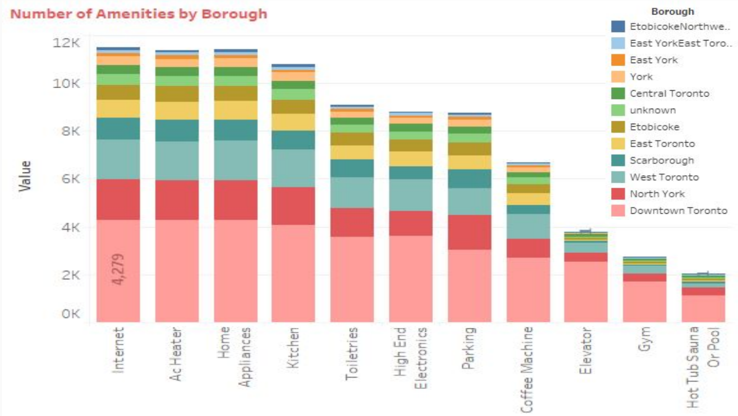




Observations:

- Majority property type - Apartment
- Most expensive property type - House
- Most expensive location - Downtown Toronto

Exploration Data Analysis



Observations:

- Highest price by accommodation - 14 guests (average price: CAD $570)
- Most popular amenity - Internet
- Hosts charge higher prices for popular amenities

Project Analysis Phase

During the last segment the team was able to uncover some key insights on how different variables can affect average Airbnb prices in Toronto. The team analyzed the data by host, location, room type, property type, average customer reviews and rating, amenities offered, etc. On average we were able to see that there is a considerable difference in average prices based on the presence of each of these factors. Currently, the team is working towards analyzing the data through visualizations to derive any insights that can help us understand the impact different factors have on average price.

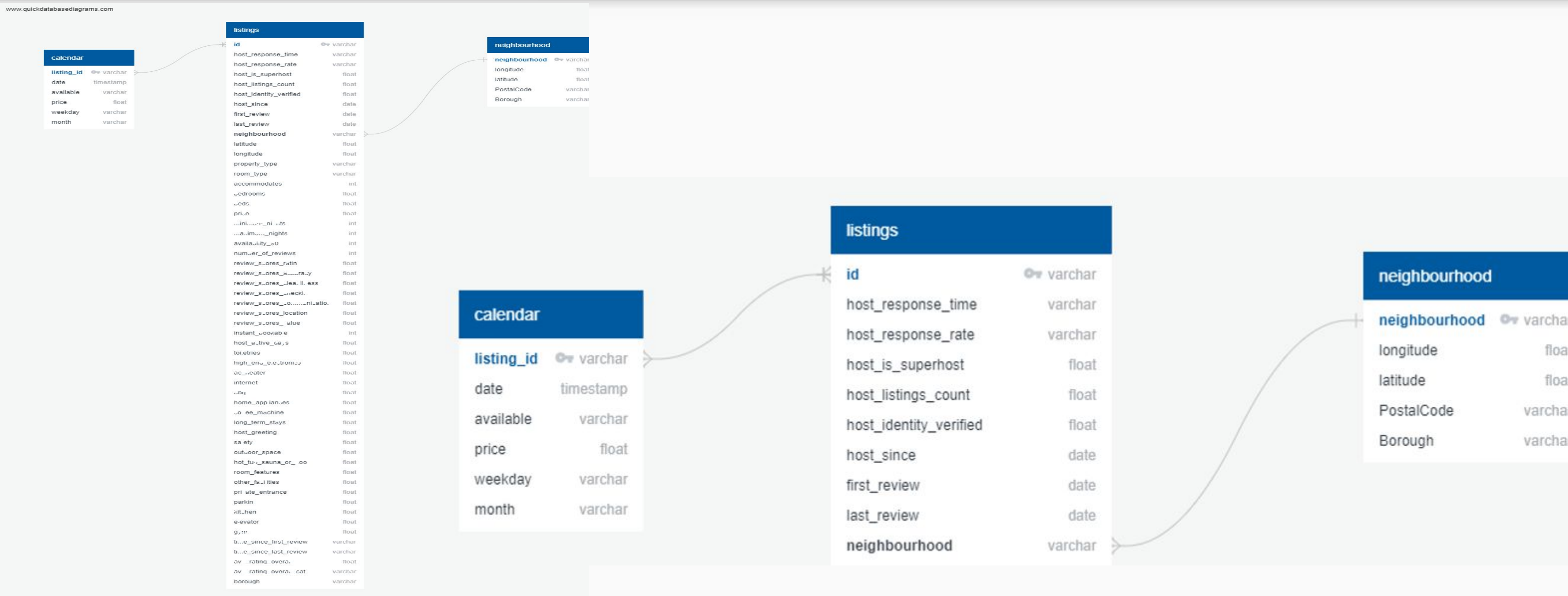
Research Questions

1. Relation between price and Room Type
2. Relation between price and Property Type
3. Top five most popular amenities
4. Top five most expensive locations
5. Relation between price and amenities
6. Relation between price and location
7. Relation between price and customer reviews and ratings
8. Popular properties by number of reviews
9. Which month has the most bookings
10. Highest number of listings by boroughs
11. Top five expensive and least expensive neighbourhoods and boroughs
12. Relation between price and host response time
13. Relation between price and host response rate

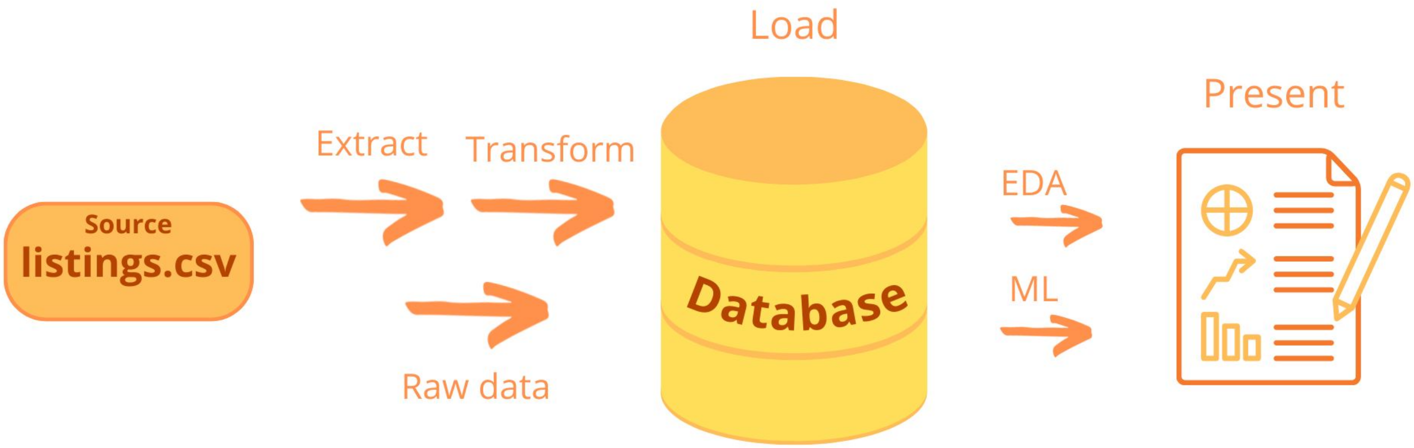
Tools Used

1. Github (Project link)
2. Database - PostgreSQL (Database details link) and Amazon Web Services RDS
3. Tableau (Dashboard link)
4. Python (Pandas, Matplotlib, Seaborn, Scikit-learn, Numpy)
5. Flask
6. HTML
7. CSS
8. Jupyter Notebook
9. Google Colab
10. XGBoost Regression Model (Scikit-learn)
11. GridSearch CV hyper-parameter tuning

ERD Diagram


www.quickdatabasediagrams.com
calendar
listing_id varchar
date timestamp
available varchar
price float
weekday varchar
month varchar
listings
id varchar
host_response_time varchar
host_response_rate varchar
host_is_superhost float
host_listings_count float
host_identity_verified float
host_since date
first_review date
last_review date
neighbourhood varchar
neighbourhood
neighbourhood varchar
longitude float
latitude float
PostalCode varchar
Borough varchar

ETL Pipeline



Machine Learning Model Results - XGBoost

Model	RMSE	R^2	RMSE(%)	R^2 (%)
Linear Regression	79.39	0.550		
Support Vector Regression	100.08	0.284	-26%	-48%
GradientBoostingRegressor	69.52	0.655	12%	19%
Random Forest Regressor	69.17	0.658	13%	20%
Extra Trees Regressor	68.77	0.662	13%	20%
XGBoost Regressor	68.26	0.667	14%	21%
XGBoost Regressor + gridSearchCV	**67.35**	**0.676**	**15%**	**23%**

- Important features:
  - Neighbourhood/Borough
  - Property type
  - Accommodates
  - Bathroom
  - Minimum nights

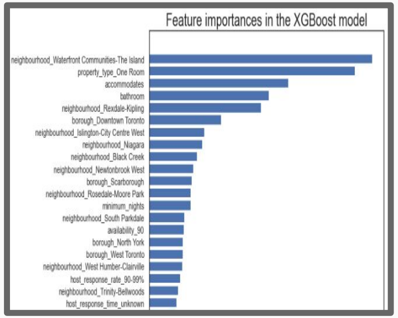
XGBoostRegressor - Hyper-parameter Tuning with GridSearchCV

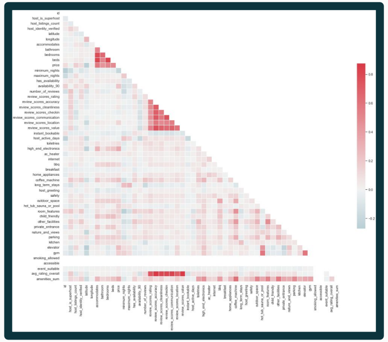
Final Model After Hyper-parameter Tuning

```python
# # Test optimum values for param_grid
param_grid = {
    'n_estimators': [100,250,500,1000],
    'eta':[0.05,0.1,0.2]
}

grid_search(param_grid,best_param)
```

```python
# # Run the XGBoost Regression model with optimized parameters
model=xgb.XGBRegressor(random_state=0, verbosity=1,**best_param)
model.fit(X_train_scaled, y_train)
training_pred=model.predict(X_train_scaled)
predictions=model.predict(X_test_scaled)
r2_score(y_test,predictions)

rmse_training=np.sqrt(mean_squared_error(y_train,training_pred))
rmse_model=np.sqrt(mean_squared_error(y_test, predictions))
print('RMSE train: %.3f' % rmse_training)
print('RMSE test: %.3f' % rmse_model)
print('R^2 train: %.3f' % (r2_score(y_train,training_pred )))
print('R^2 test: %.3f' % (r2_score(y_test, predictions)))
```

```
RMSE train: 37.916
RMSE test: 67.563
R^2 train: 0.901
R^2 test: 0.672
```





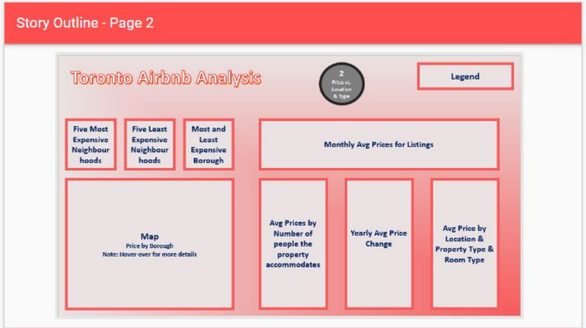
Results:

- RMSE test score: 67.35
- R-Squared score: 67.6%
- Important features:
  - Neighbourhood
  - Property type
  - Accommodates
  - Bathroom
  - Borough
  - Minimum nights
  - Host response rate

Story Outline

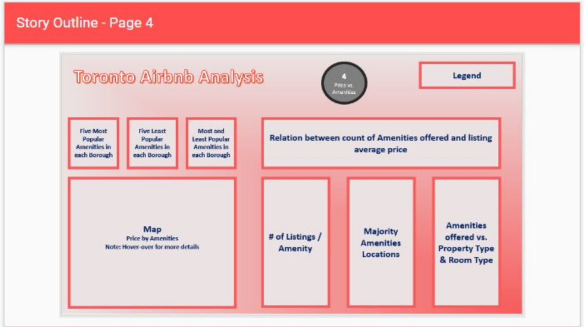
Page 1

Page 2

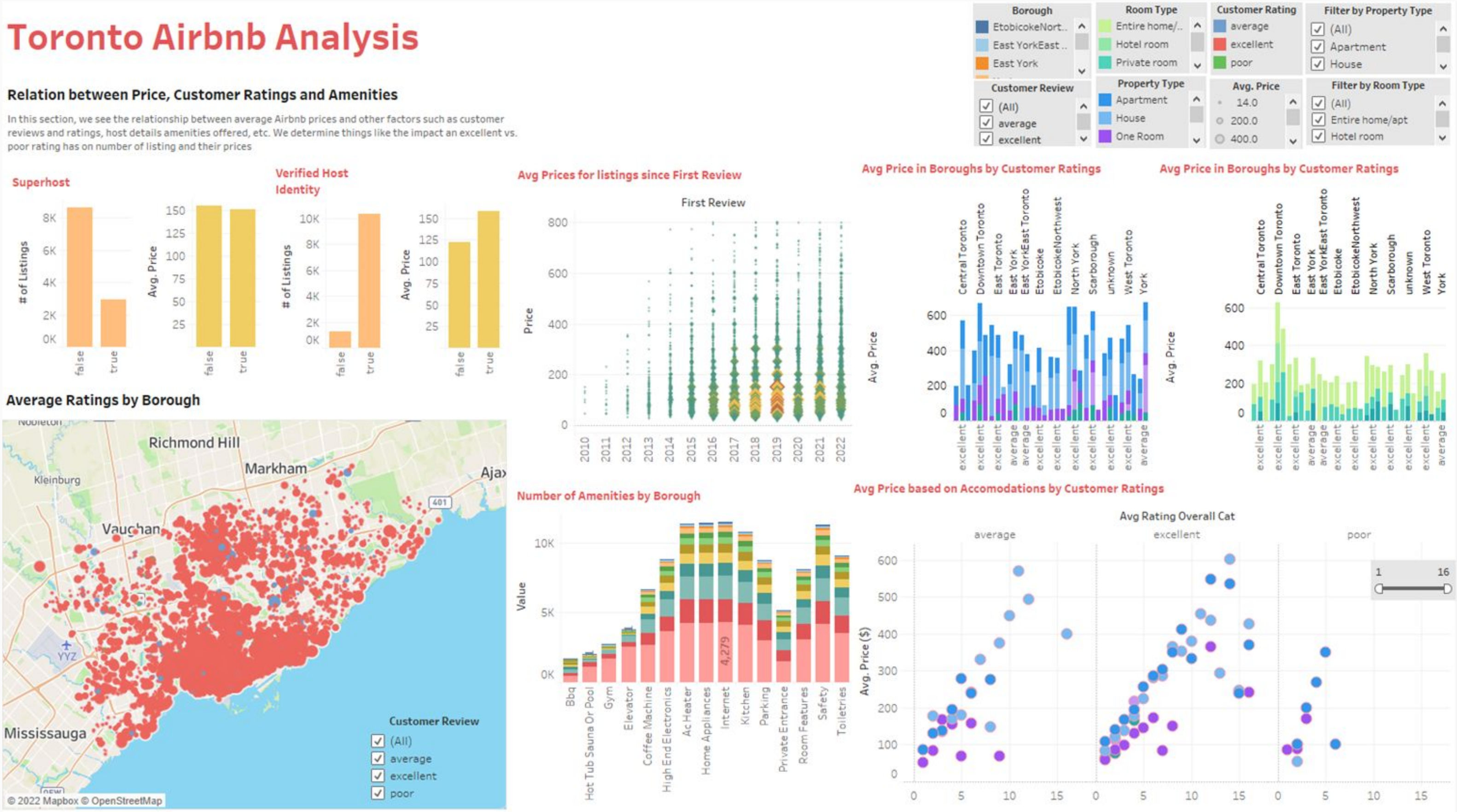


Page 3

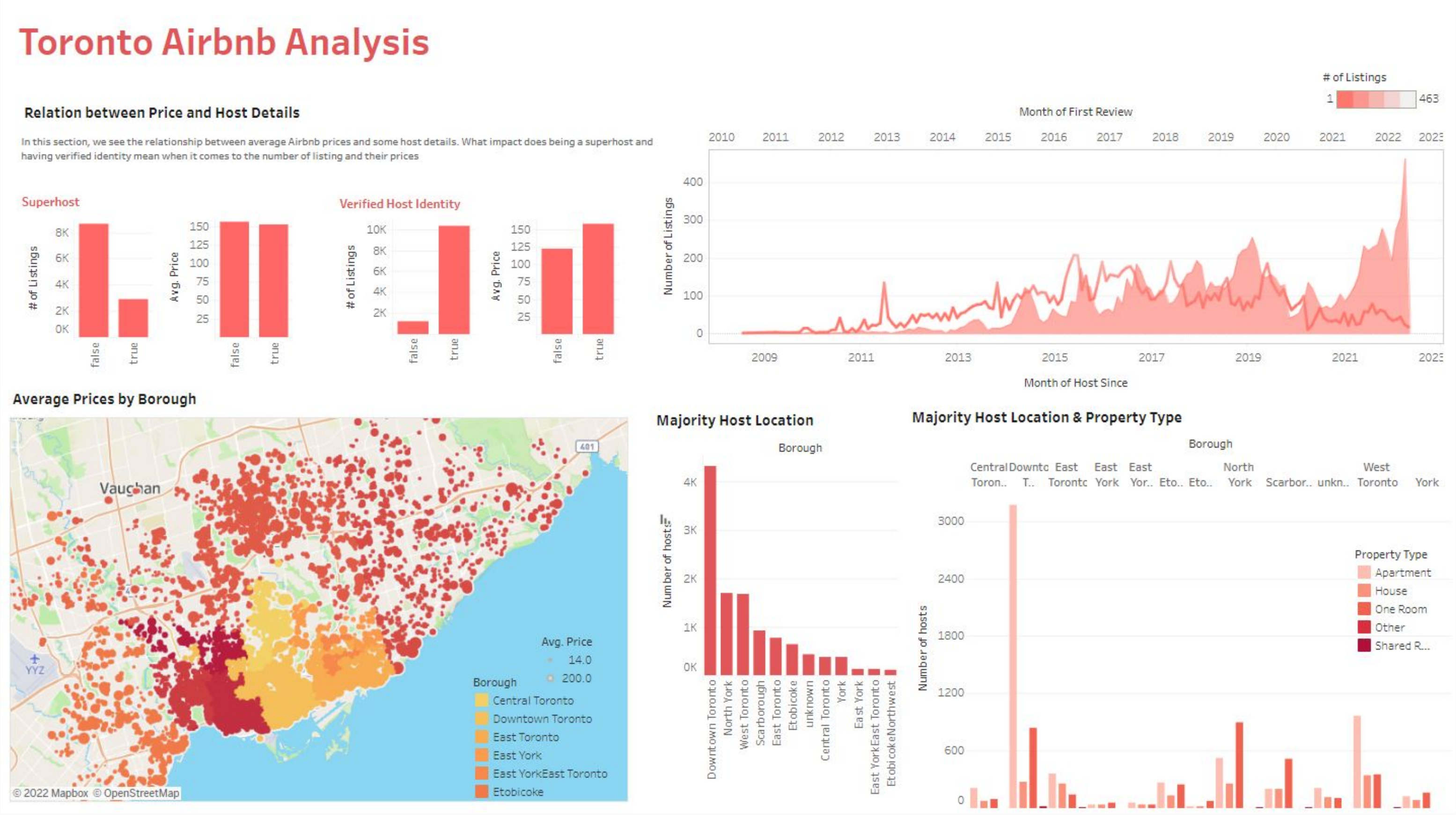
Page 4



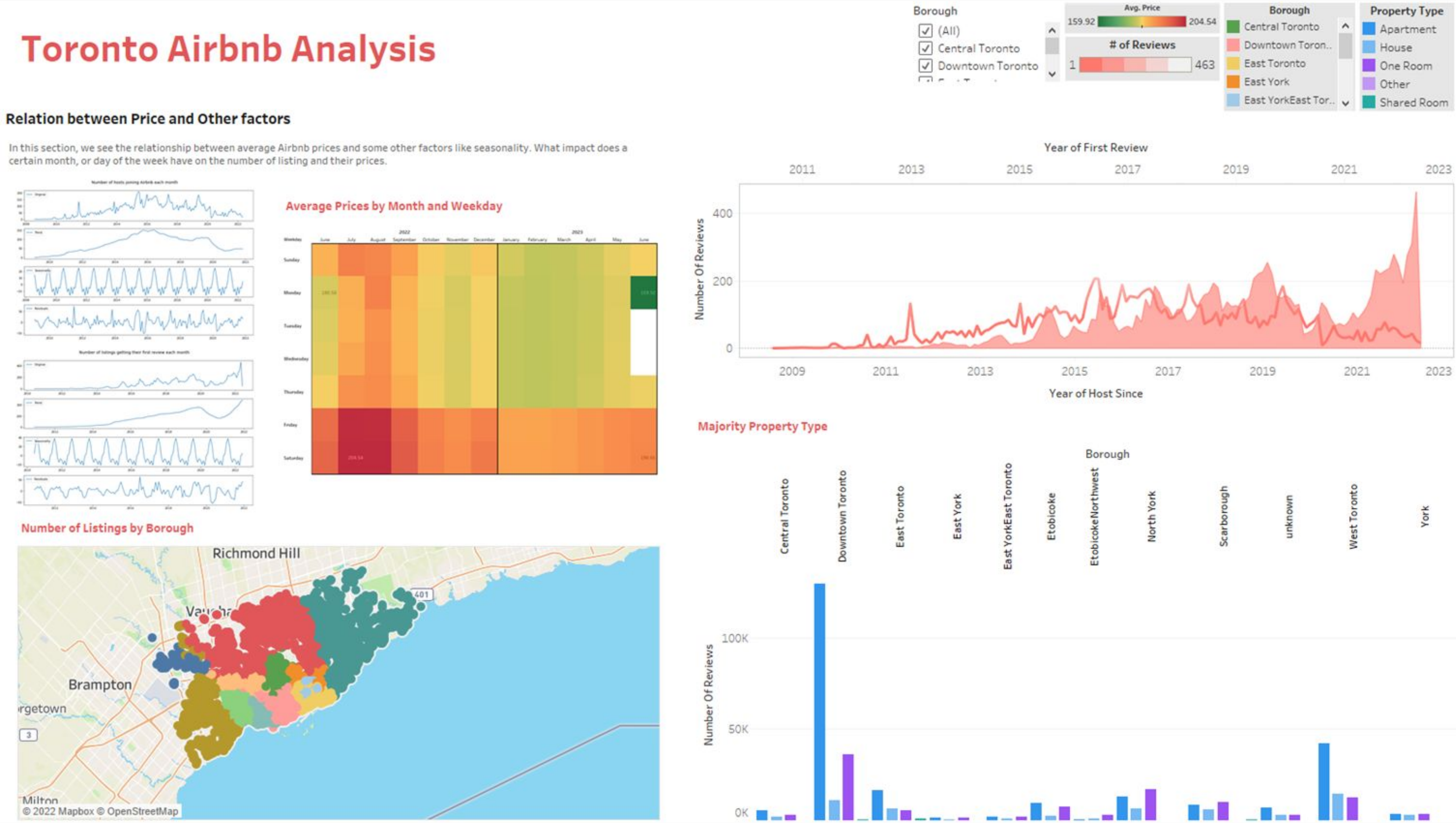
Final Dashboard



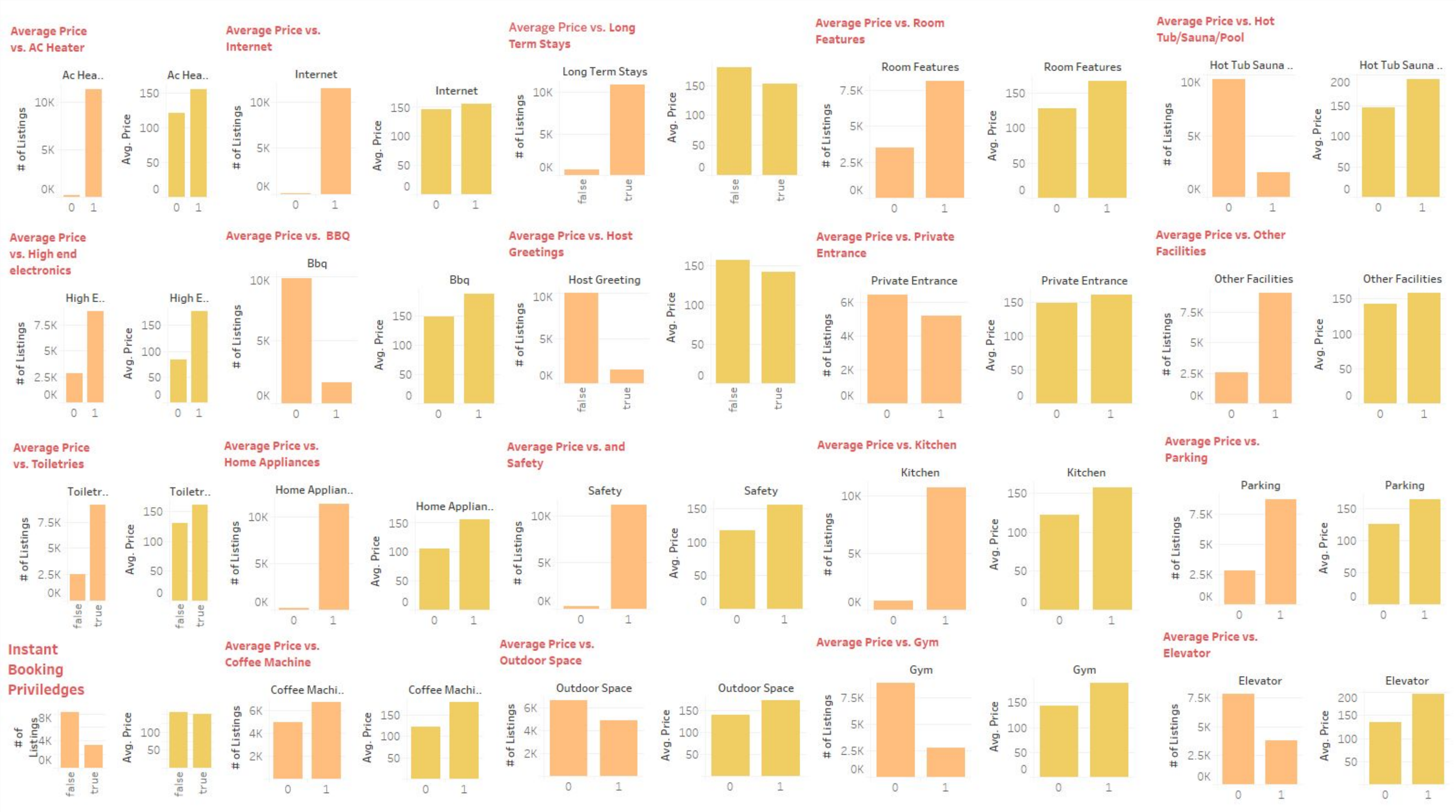
Final Dashboard



Final Dashboard

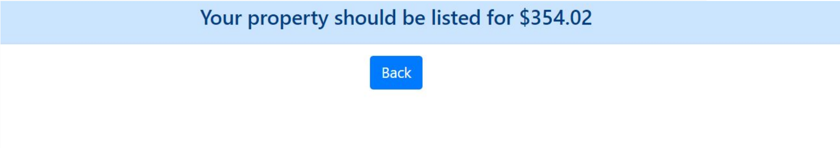


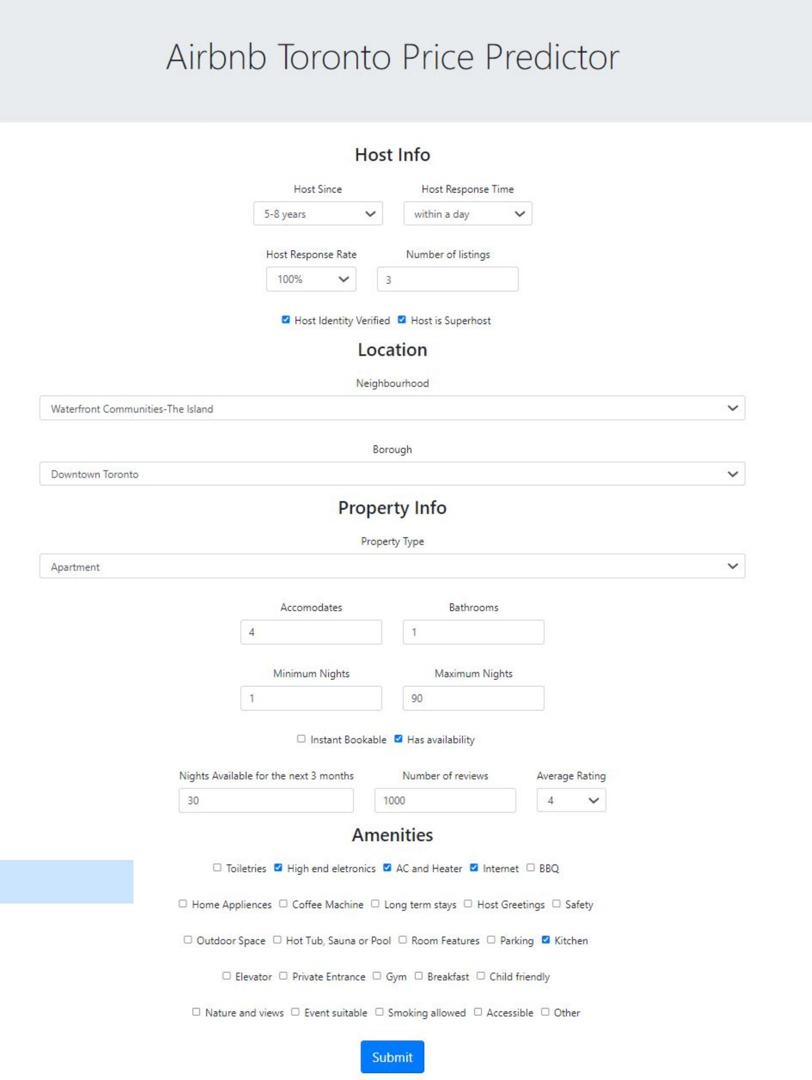
Price vs. Amenities



Webapp

We created an exported machine learning model on webapp with Flask and HTML. This webapp can be used by potential Airbnb hosts to predict optimal prices for their properties.





Limitations and Future Improvements

- Limited data volume
- Lack of data on important factors like bookings, cancellation policy, security deposit, etc.
- No data regarding points of interests, restaurants or cafes around the property
- Have not performed sentiment analysis on customer reviews and ratings