Project Report

Airbnb Big Data Mining

Team: LZY - Brian Zhou, Hongying Yue , Yiran Lu

Dec 7, 2023

Problem Definition

Inspiration

Airbnb, as a prime example of the sharing economy, provides a platform that facilitates short-term lodging rentals and related travel experiences, with well known for its data-driven approaches.

We were attracted by its success and generated the idea to analyze how the rental service of Airbnb is making differences on the regular long-term rental market of Vancouver.

Dataset

After searching for possible data to implement the analysis, the follow two datasets are our focus:

- Inside Airbnb: detailed Airbnb listing data, including quantitative customer evaluation scores to each listing
- CMHC Housing Market data: latest housing market data of Canada, including rental price data down to neighbourhood in Vancouver
- REGBV (Real Estate Board of Greater Vancouver): Vancouver neighbourhood definitions with geographical data

Objective

Besides our major interested question, we would like to have the project's findings to contribute valuable insights for both property owners and renters.

Therefore, we defined concrete questions to answer:

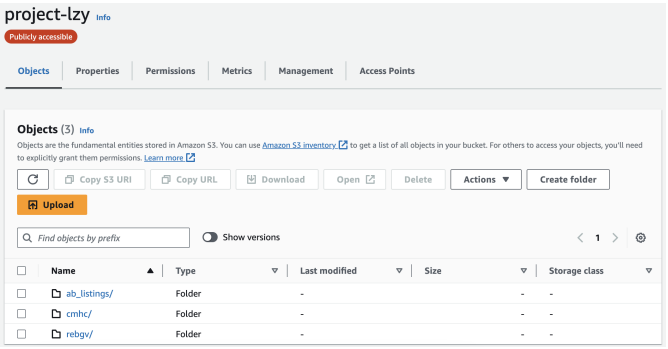
- **In which season of the year can you get the best rental prices?**
- **Does Airbnb offering in a neighbourhood affect long-term rental prices of the region?**
- **What factor determines the attractiveness of an Airbnb listing?**

With the answers to those questions, renter customers can make informed travel plans and rental decisions, and Airbnb hosts can provide better and competitive services.

Methodology

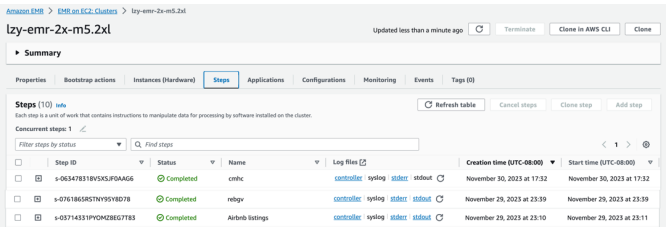
In order to practice what we learnt from the class and try out relevant new tools, we chose specific applications accordingly.

1. AWS S3 - Data storage



As learnt from the assignments, in order to make preparation for using AWS EMR Spark to conduct ETL, we chose AWS S3 for data storage, where most EMR applications can take inputs easily. In addition, as the object storage service for AWS, S3 stands out for scalability, durability, security, versatility, ease of use, cost-effectiveness, and global reach, which will ensure the expansibility of the project in a real production environment as the raw data grow significantly.

2. AWS EMR Spark - ETL



We applied AWS EMR Spark as the major tool for ETL. Besides its popularity among external organizations, EMR has plentiful integrations with other AWS services, which provides great capability for further data mining and analysis. Of course, we could enhance the skills we learnt from the course, and EMR offers a ready-made environment to run PySpark conveniently.

3. Spark ML - Data mining

Within the process of data analysis, apart from direct visualization, we'd like to seek for deep explanations via machine learning methods. Given a big data project implementation, Spark ML jumped out as the first choice. Based on our project proposal, a question regarding finding the attractiveness of an Airbnb listing is highly suitable for employing the approach of Assignment 10. Therefore, we followed the instructions and extracted the feature importance to draw analysis conclusion.

4. Flourish - Data visualization

When it came to the visualization of our datasets, we initially considered using D3.js to develop custom visualization templates. D3.js offered the flexibility to create highly tailored visual representations

of our data. However, after evaluation and experimentation, we opted to utilize the data visualization platform Flourish. Our decision to switch to Flourish was driven by its ability to abstract the complexities of the D3.js framework, thereby simplifying the visualization process. This made it easier for us to iterate on our visualizations and also proved to be powerful enough to meet our project's needs.

5. Jupyter Notebook - Data analysis

In order to quickly verify assumptions during analysis, we chose Jupyter Notebook as a supplementary tool. Compared to other graphing applications, Jupyter Notebook can take the advantage of Python and its libraries like MatPlotlib and Seaborn, allowing us easily convert the cleaned data to desired charts, which helps to testify if the output is consistent with our hypothesis more efficiently. Besides, it's easy to save the intermediate analysis, providing agile presentations for team communication.

Difficulties

Real world can never be as perfect as you expected. Although we knew what we wanted to excavate and thought we had powerful tools and sufficient skills for the project, we encountered several challenges while collecting data for big data analysis.

1. Data availability

A primary issue was the nature of the data availability. We often found data that was either very detailed in scope, such as rent prices detailed per neighborhood, or data with high frequency, like monthly reports. However, these datasets frequently lacked in the other aspect; detailed datasets were often only available as annual reports, while high-frequency datasets typically included only broad pricing information per zone. This dichotomy in data availability posed significant challenges in achieving a comprehensive analysis.

2. Data accessibility

Another significant hurdle was the accessibility of datasets. Many dataset providers or agencies tended to restrict access to detailed reports behind paywalls. Moreover, there was a common practice of providing access only to the latest datasets, with no availability of historical data. This limitation greatly impeded our ability to meet several of our project's goals. Despite these challenges, we remain optimistic that, over time, we can accumulate sufficient data to conduct more in-depth analyses.

3. Format inconsistency

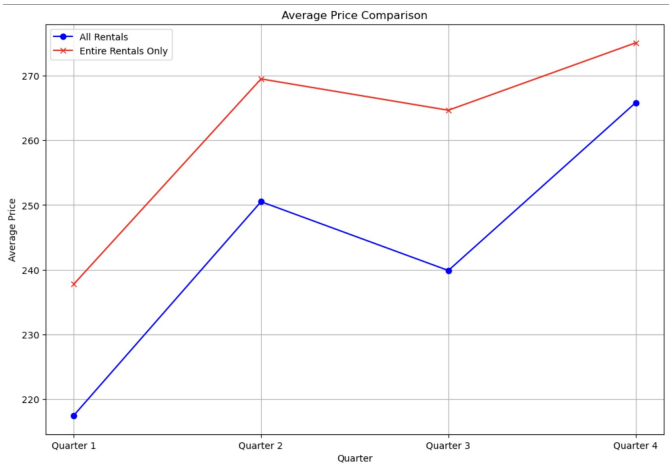
The format of the datasets presented another layer of complexity. Many datasets were available in less convenient formats, such as PDFs or even images. To address this, we utilized OCR to extract text and data for processing in Spark. This approach, however, introduced its own set of challenges. Automating the ETL process was difficult as it often required manual corrections post-OCR, affecting the efficiency and scalability of our data processing pipeline.

Another issue we faced was the inconsistent naming format for different neighborhoods. We ended up manually creating reference tables for different names in order to successfully process all the datasets.

Results

Our project helps us to find out the answers to our interested questions.

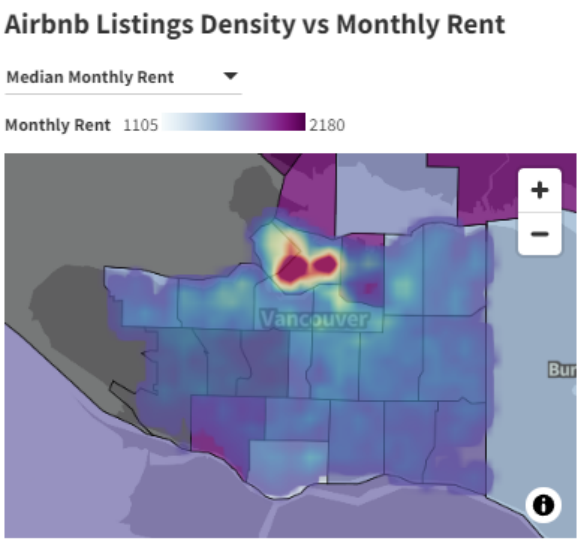
Season to get the best rental offer



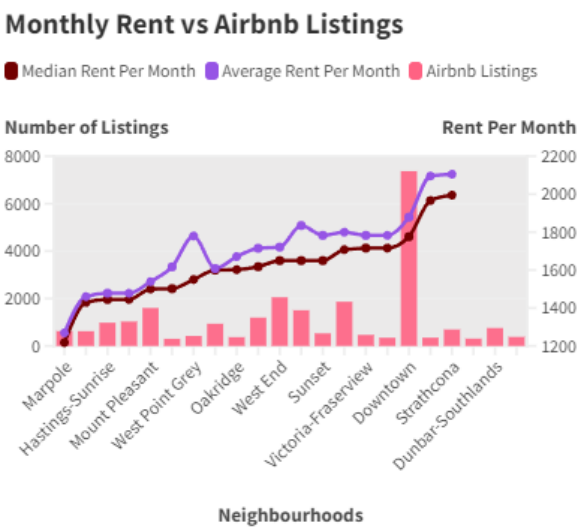
We plot the average rent price for each quarter and we can see that the 4th quarter has the highest price listing. We also plot the average price for entire rental only for each quarter, just to see what would the difference be compared to all Airbnb rentals (for example single rooms).

In conclusion, the rent would be much higher in later of the year and entire rental would have a higher rental price compared to all rentals, but their trend is the same.

Airbnb's impact on local rent prices



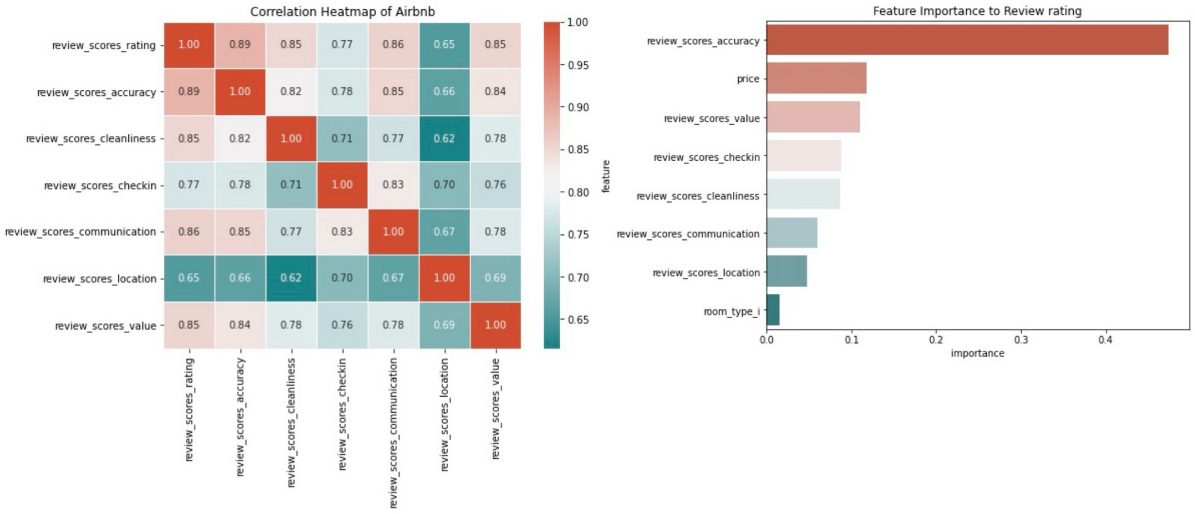
Source: CMHC, Inside Airbnb



Source: CMHC, Inside Airbnb

With the limited datasets we have on hand, we concluded that Airbnb's density doesn't have a profound impact on local rental pricing. Although we can not verify the relationship, this index is still worth tracking in the future.

Determining factor of the attractiveness of an Airbnb listing



To find the answer to this question based on our data at hand, we calculated the correlation coefficients between the review rating with 6 detailed evaluated factors. In addition, we built a regression model with GBTRegressor from those 6 variables together with additional 2 potential factors of price and room type. The tree model provided feature importance of all the 8 factors, which serves as a second source to draw a conclusion.

Combining the results of these two steps, we found that the accuracy of a listing description stands out as the most influential factor, contributing a substantial importance of 47.5%. A suggestion to an Airbnb host could be, in spite of less attractive room, being honest to customers can not go wrong.

Takeaways from implementation

The main takeaway we learnt from the implementation is that the data cleaning is as important as actually analyzing the data. We encountered a lot of issue while cleaning the data, like the format of the data, ways of collections, etc, and we could only produce valuable and useful information after we cleaned and output the data in the format we need. People usually value more on the analysis part, but it is cleaned data that make everything possible.

Project Summary

Give yourself a total of 20 point in these categories:

Table 1: Project Summary

NO	Category	Score
1	Getting the data	3
2	ETL	5
3	Problem	3
4	Algorithmic work	3
5	Bigness/parallelization	2
6	UI	1
7	Visualization	2
8	Technologies	1