Sonder Data Analysis Project

Realised by

Suibgui Khaireddine

Academic Year: 2023/2024

Table of Contents

List of Figures

List of Tables

General Introduction

Business Intelligence is a rapidly evolving topic that encompasses both management and business functions. It serves as a decision tool that enables understanding of the overall activities and vision of an organization.

This view of the business environment requires a deep understanding of various business functions and specific organizational specifications. The implementation of Business Intelligence projects cannot be undertaken without defining a comprehensive BI strategy. Therefore, we can say that Business Intelligence is the process of extracting information and knowledge from data.

Today, the BI market offers comprehensive solutions through analysis of reporting. The key to successful analysis is obtaining information from data, which facilitates better decision-making.

Business Intelligence addresses business challenges by transforming large volumes of data into actionable insights that can drive business objectives and achieve revenue goals. Business insights can be delivered through various data presentations such as reports, dashboards, and visualizations.

Generally, there are three different ways to integrate business intelligence into an organization:

- Managed reports that are periodically refreshed.
- Self-service analytics.
- Inputs for operational systems.

By leveraging these approaches, businesses can harness the power of data to gain meaningful insights and make informed decisions to drive their success.

Chapter I: General Framework of the Project

I.1 Project overview:

I.1.1 Problematic

Previously, the concept of booking for tourists, whether national or international, was a bit complicated in terms of money.

- **Price** is still the main constraint for tourists. Reservations have become increasingly expensive, especially in hotels.
- **The type of booking** is also important, such as half-board, full-board or all-inclusive. In general, it's the hotels that demand all the rules for a trip, and the booker is the aggrieved party.
- In **the absence of democratization**, we've fallen into the problem where the rich can travel all over the world and the poor can't even spend a holiday in their own country.

I.1.2 Proposed solution

Globally, the price is the first constraint for a traveler, so, we're going to deal with price prediction based on several criteria.

Our solution is based on:

- Data collection.
- Machine Learning algorithms.
- Dashboarding on a user interface.

Finally, we'll look at the concept from an economic and commercial point of view.

I.2 Work Methodology

The CRISP method breaks down into 6 stages, from understanding the business problem to deployment and production.

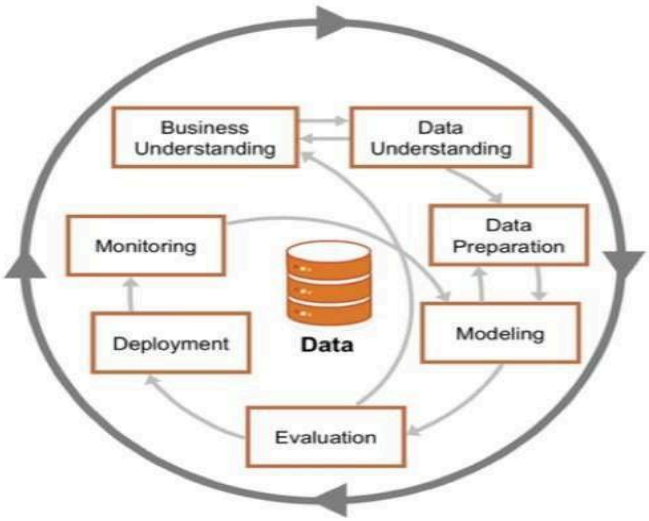

Figure 1 : CRISP

The CRISP program comprises 6 phases:

1 - Understanding the business.

2 - Determining business objectives.

3 - Assessing the situation.

4 - Determine data science objectives.

5 - Produce project plan

- Collect initial data
- Describe data
- Explore data
- Check data quality
- Select data
- Clean data
- Select the modeling
- Build model
- Evaluate model
- Evaluate results

6 - Plan deployment

- Plan follow-up and maintenance
- Product final report

I.3 conclusion

In this first chapter, we identified the problems by giving adequate solutions with a presentation of the adopted methodology. In the next chapter, we describe in detail the proposed prediction analysis system.

Chapter II: Price prediction system

II.1 Introduction

After presenting the general context of our project in the previous chapter, we devote this chapter to the theoretical study of our proposed Sonder rental price prediction system.

II.2 Rental Price prediction system architecture

The system we propose consists of two main phases:

The first step is to discriminate between the data in our database. This stage comprises three sub-steps: pre-processing, data cleaning and regression. Prediction is based on Machine Learning (ML) algorithms.

The second step is to visualize the results in an interface, using Power BI.

II.3 BI data processing

This section is divided into two subsections. In the first subsection, we give a general introduction to the field of Business Intelligence. The second part is devoted to the Dashboard.

II.3.1 Business Intelligence

Business Intelligence is a set of processes, architectures, and technologies for converting raw data into meaningful information that leads to business actions.

It is a suite of software and services that transforms data into actionable information and insights that have a direct impact on strategic, tactical, and operational business decisions.

BI tools perform data analysis and create reports, summaries, dashboards, maps, charts, and graphs to provide users with detailed information about the nature of the business.

Figure 2 : Business Intelligence

II.3.2 Microsoft Power Bi

Power BI is a set of software services, applications and connectors that work together to transform disparate data sources into immersive and interactive visual information.

Figure 3 : Power BI

II.4 Price prediction process

In this section, we outline the basic concepts used in our solution. Next, we present the architecture of the price prediction system.

II.4.1 The main concepts used in our solution

Before presenting our solution, we'll first explain the main phases of our system and the techniques used.

II.4.1.1 Data base:

Feature	Data
Type	.CSV
Size	6.50 Mo
Number of attributes	16
Number of observations	48895

Tableau 1 : Database

II.4.1.2 Data type

- **Quantitative Data:**

Attribute	Description
Id	Identificator
Host_id	Identificator of Host
Latitude	Latitude
Longitude	Longitude
Price	Price
Minimum_nights	Minimum nights

Number_of_reviews	Number of reviews
Reviews_per_month	Reviews per month
Calculated_host_listings_count	Calculated host listings count
Availability_365	Availability 365
Last_review	Last_review

Tableau 2 : Quantitative Data

❖ **Qualitative Data:**

Attribute	**Description**
Neighbourhood_group	Neighbourhood Group
Neighbourhood	Neighbourhood
Room_type	Room Type
Name	Name
Host_name	Host Name

Tableau 3 : Qualitative Data

II.4.1.3 Techniques used

In this section, we present the techniques used in our project.

- **Machine learning:** ML is an application of artificial intelligence. AI gives systems the ability to learn and improve automatically from experience without being explicitly programmed.
- **Supervised learning:** The learning algorithm receives labeled data (number of classes known in advance) and the desired output.

II.4.2 Prediction process steps

During this section, we will present the steps involved in implementing our process. In fact, Collaboratory is a cloud service offered by Google, based on Jupyter Notebook, and designed for training in machine learning. This platform enables Machine Learning models to be trained directly in the cloud. This means we don't need to install anything other than a browser on our computer.

II.4.2.1 Data acquisition

The first step is to identify the sources of data to be used. These data may be qualitative or quantitative, supervised or unsupervised. Our database contains supervised qualitative and quantitative data.

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355

Figure 5 : Dataset

II.4.2.2 Data Preparation

Once the data has been successfully extracted, we go through the following pre-processing process:

- Remove duplicate rows.
- Remove attributes not required in the process.
- Fill in missing values.
- Remove attributes after correlation calculation.

II.4.3 Regression

In this part, we study machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, KNN.

II.4.3.1 Linear Regression

Linear regression analysis is used to predict the value of one variable as a function of the value of another variable. The variable whose value you wish to predict is the dependent one. The variable you use to predict the value of the other variable is the independent one.

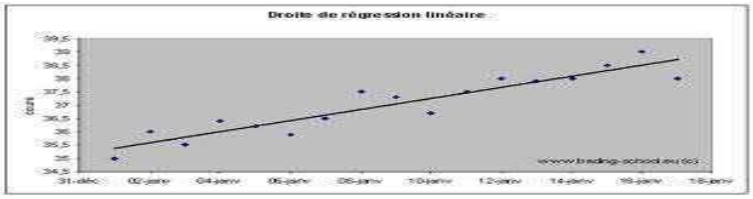

Figure 6 : Linear Regression

II.4.3.2 Decision Tree

Decision trees can be used to predict an actual quantity (for example, the price of a room in a location), in which case the prediction is a numerical value.

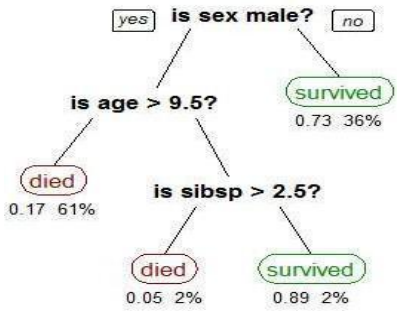

Figure 7 : Decision Tree

II.4.3.3 KNN

The K-NN (K-nearest neighbors) algorithm is a supervised learning method. It can be used for both regression and classification. Its operation can be linked to the following analogy: "Tell me who your neighbors are, and I'll tell you who you are".

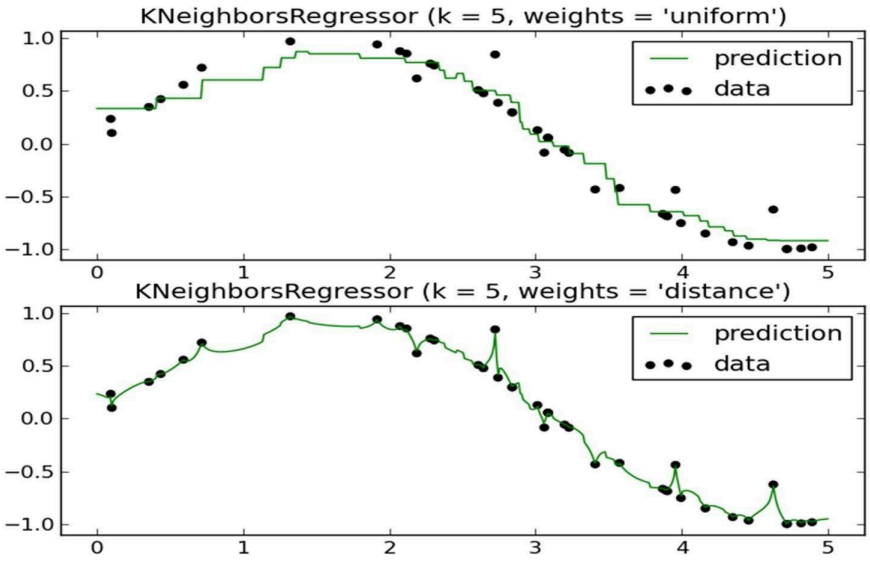

Figure 8 : KNN

II.4.3.4 SVR

A version of SVM for regression was proposed in 1996. This method is called support vector regression (SVR). The model produced by support vector classification depends on only a subset of the training data, as the cost function for model building does not care about training points that lie beyond the margin.

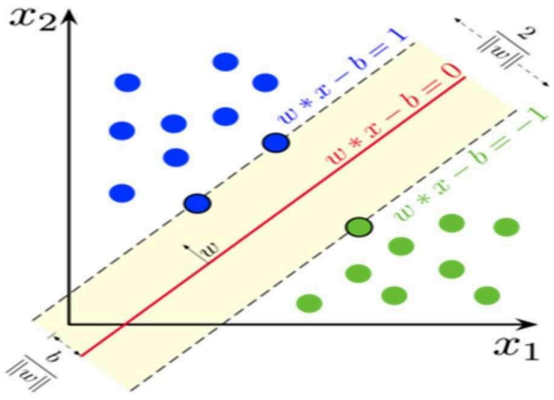

Figure 9 : SVR

II.5 Visualization

This step involves determining the right data outputs in terms of the quality of price predictions. There are a number of applications with a dashboard displaying useful information for the process, and in our case, we'll be using Microsoft Power BI.

Figure 10 : Power BI

II.6 Conclusion

In this chapter, we have presented the overall architecture of our process, which is divided into two main phases: regression and visualization. The first process consists of two sub-steps: pre-processing and cleaning.

The second is the visualization of the results in a Dashboard. In the next chapter, we'll take a closer look at the implementation of these models, analyze the results in order to select the best-performing model.

Chapter III: Realization of prediction systems

III.1 Hardware environment

In this section, we present the hardware environment. Our system is implemented on a "DELL" computer with these characteristics:

Processor : Intel Core i5

RAM : 12 GO

Hard Disk Drive : 500 GO

OS : Windows 11

III.2 Dashboard

In this step, I tried to create a dashboard on Microsoft Power Bi for simple data visualization.

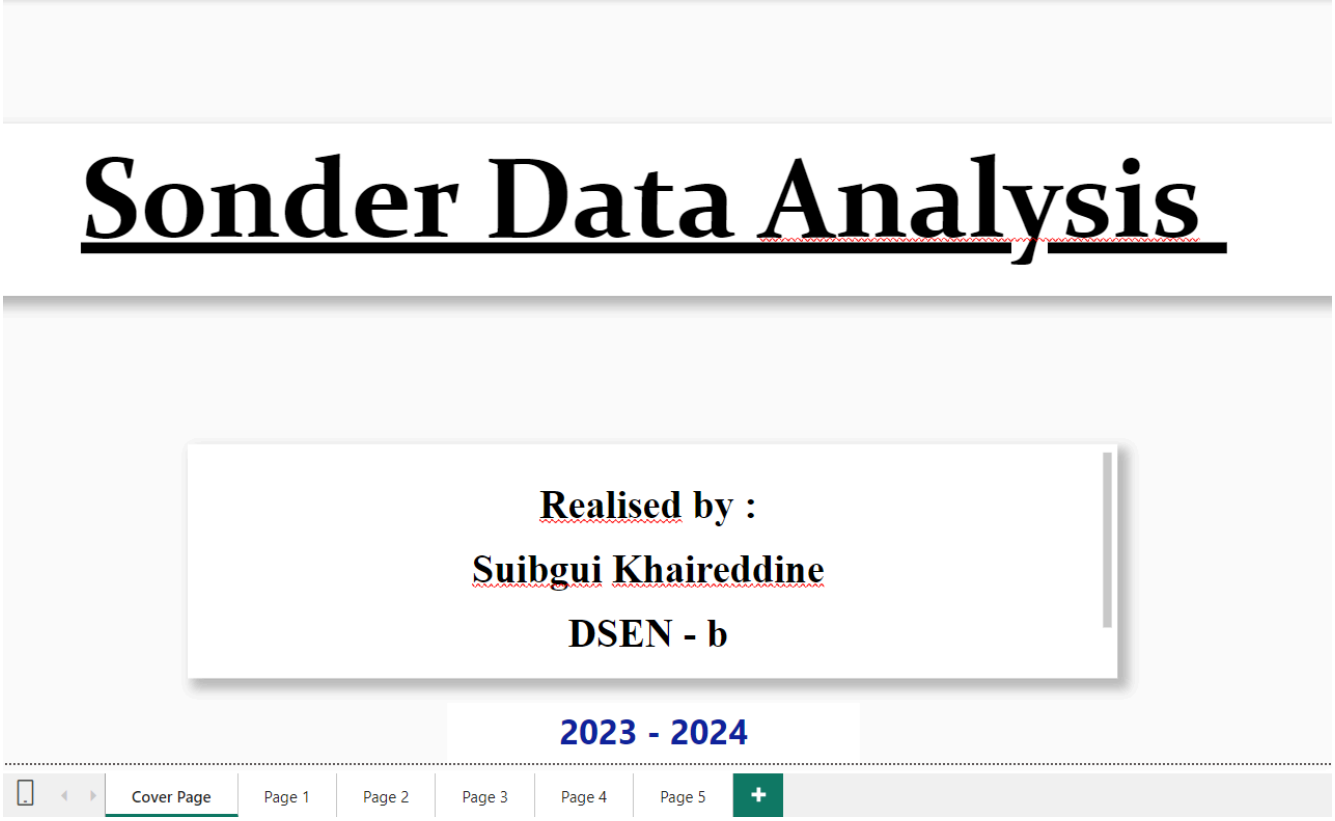

Figure 11 : Dash 1

Figure 12 : Dash 2

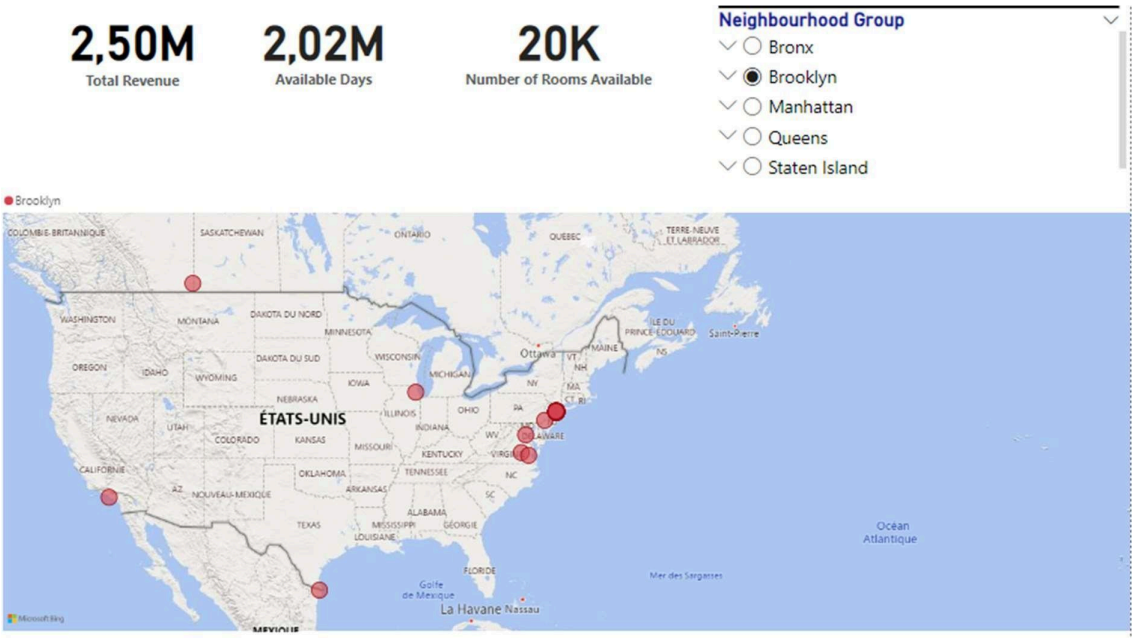

Figure 13 : Dash 3

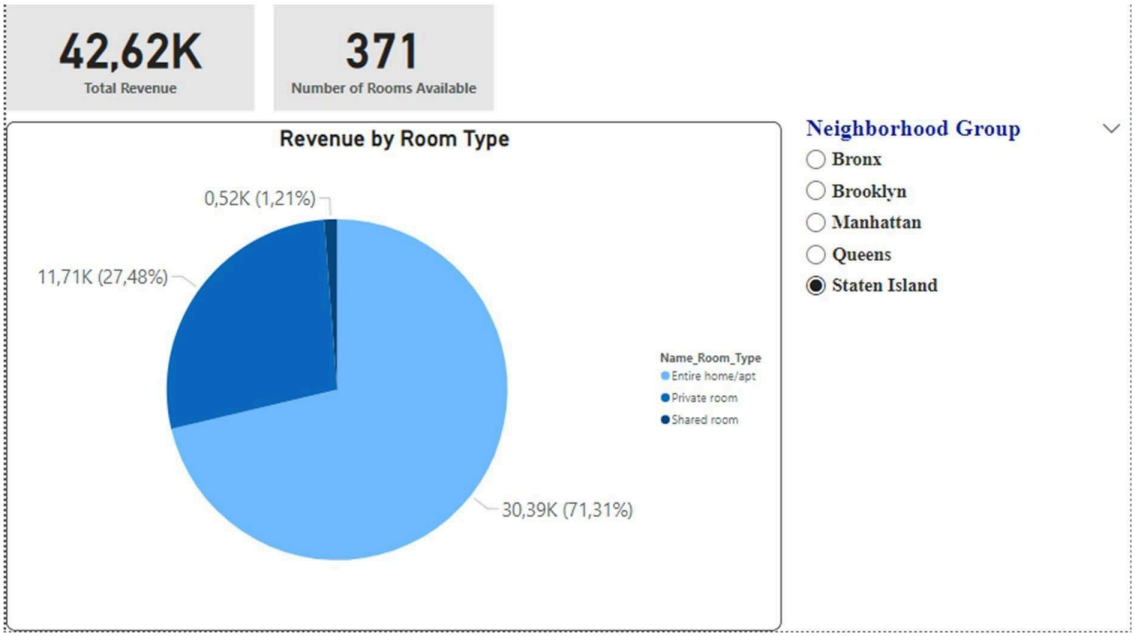

Figure 14 : Dash 4

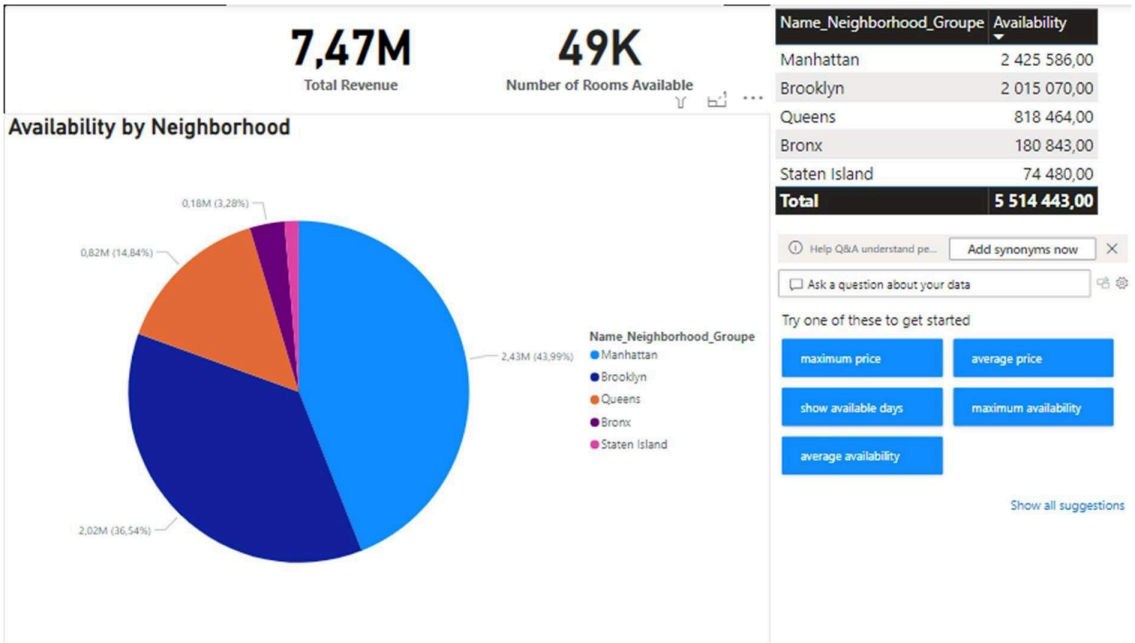

Figure 15 : Dash 5

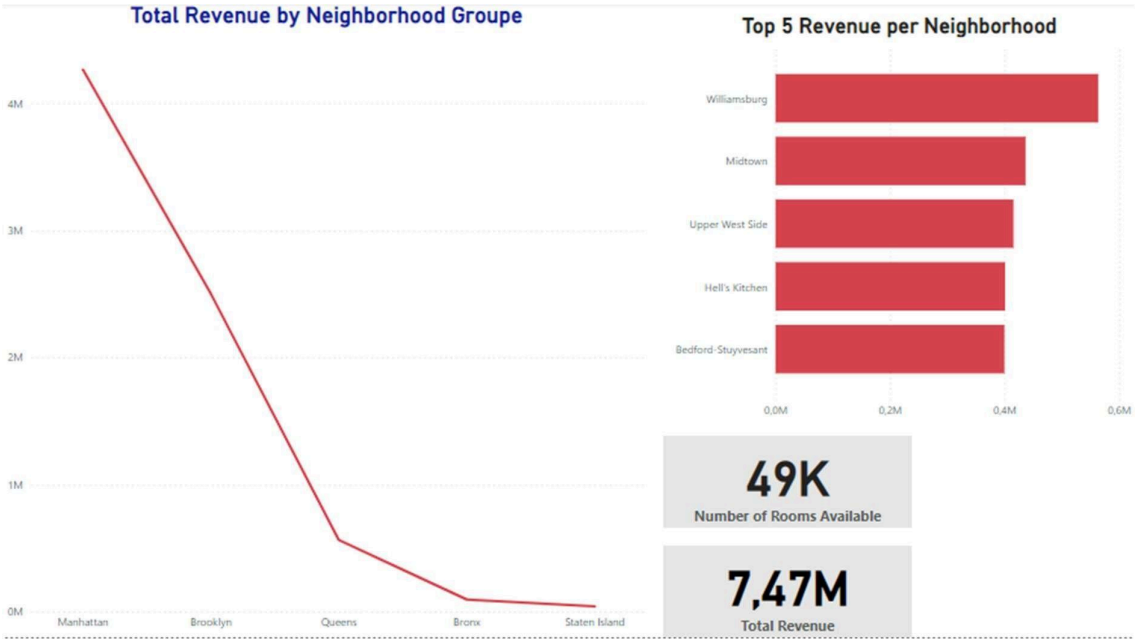

Figure 16 : Dash 6

III.3 Machine Learning results

III.3.1 Library used

- NumPy: is a library for the Python programming language, designed to manipulate matrices or multidimensional arrays.
- Pandas: is a library for data manipulation and analysis.
- Matplotlib: is a Python programming language library for plotting and visualizing data in graphical form.
- Seaborn: is a library for creating statistical graphs. It is based on matplotlib.
- Sklearn: is a free machine-learning library for Python. It features various algorithms such as support vector machine, random forests, and k-neighbors, and also supports Python numerical and scientific libraries such as NumPy.

III.3.2 Experimental Protocols

In this section, we measure the performance of our algorithms, a vital important step in producing a quality algorithm that meets business expectations.

In our case, we consider these metrics:

- **Root Mean Square Error (RMSE) [0, inf]:**

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_r - y_p)^2}$$

Figure 19 : RMSE

- **Explained Variance (EV) [0,1] :**

$$\mathrm{EV} = 1 - \frac{\mathrm{var}(y_r - y_p)}{\mathrm{var}(y_r)}$$

- **R^2 Statistic :**

$$R^2 = 1 - \frac{RSS}{TSS}$$

III.3.3 Data preparation

In this step, we detail the results of the cleaning and pre-treatment phases mentioned in the previous chapter.

- **Removing duplicates :**

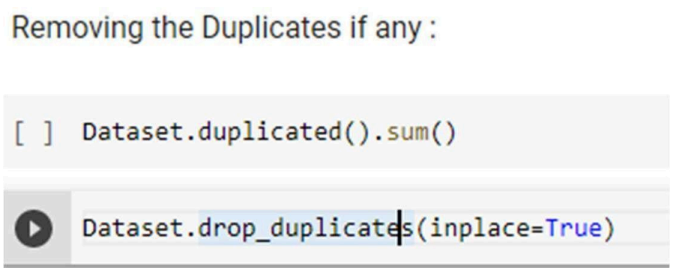

Figure 17 : Removing duplicates

- **Drop unnecessary columns :**

Figure 18 : Drop unecessary columns

- **Filling NA :**

Figure 19 : Filling NA

- **Removal of columns when calculating the correlation:**

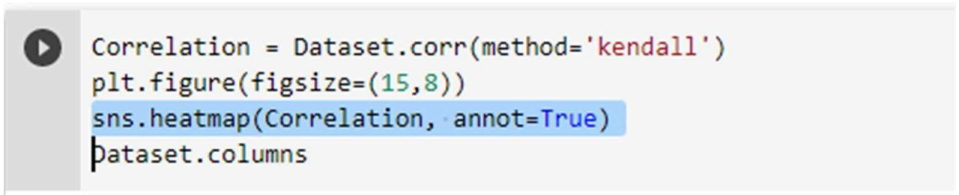

Figure 20 : Correlation

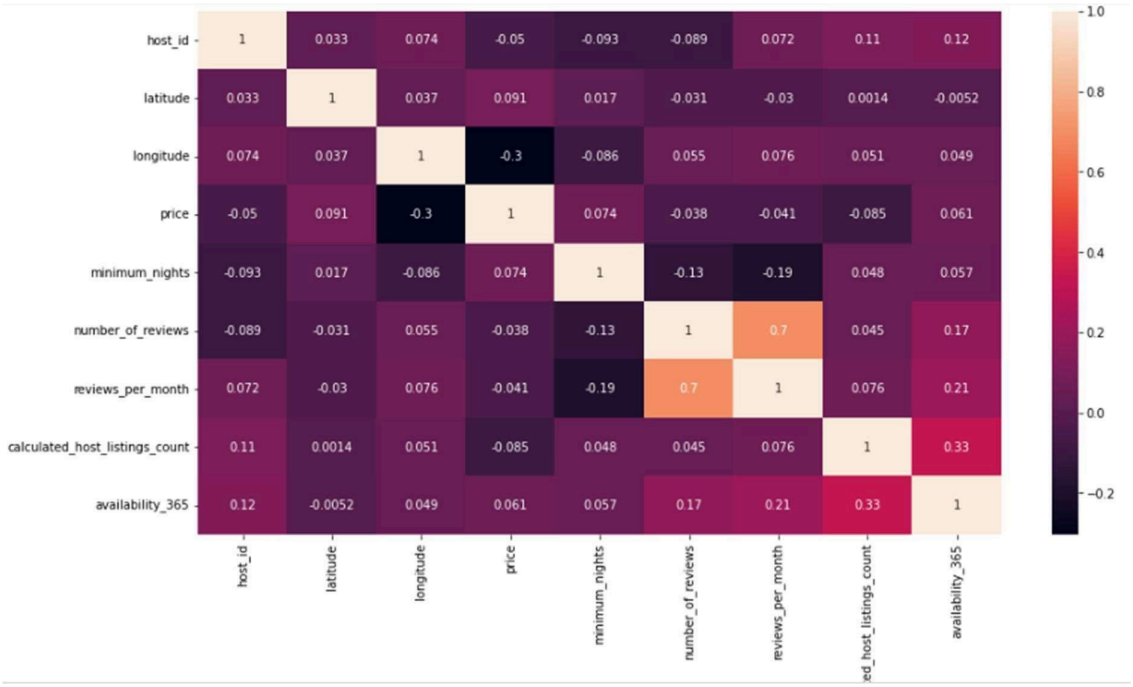

Figure 21 : Correlation Map

III.3.4 Data pre-processing

This phase details the pre-processing steps mentioned in the previous chapter, namely the conversion of categorical data.

Categorical data conversion: in this step, we'll be exploiting analysis techniques, as categorical data cannot be consumed directly by a learning model.

- **Label Encoder**

```
from sklearn.preprocessing import LabelEncoder
Dataset_cat=Dataset.drop(['price', 'minimum_nights','calculated_host_listings_count','availability_365'],axis=1)
le=LabelEncoder()
for i in Dataset_cat:
  Dataset_cat[i]=le.fit_transform(Dataset[i])
Dataset_cat.head()
```

Figure 22 : Encoder

	neighbourhood_group	latitude	longitude	room_type	reviews_per_month
0	1	2025	4568	1	21
1	2	11521	3450	0	38
2	2	16012	7576	1	0
3	1	4964	5791	0	464
4	2	15154	7367	0	10

Figure 23 : Values encoded

III.3.5 Model evaluation

As mentioned above, for the evaluation of the price prediction model, we used the MSE, R2, Explained variance and computational complexity as protocols. The table shows the results of the evaluation of the five algorithms used (Linear regression, Decision Tree, Random Forest, KNN regressor and Xgboost). From this table we can clearly see that KNN regressor has the best performance in terms of variance explained, R2 statistic, computational complexity with k=1, as it is generally most dedicated to numerical data.

In terms of comparison between Random Forest, Decision Tree and Linear regression, we can clearly see that Random Forest is the best in terms of variance explained, R2 statistic, as it is a good classifier of the bagging category.

In terms of computational complexity, we can see that KNN is the best performer with the shortest runtime, followed by Decision Tree, Random Forest and Xgboost with long runtimes. Lastly, Xgboost has a very long execution time of around 32s to process and analyze the data. This computational complexity is due to the use of multiple layers, so data learning is performed incrementally. Our choice is KNN, as it gives the best results in terms of computational complexity, variance explained, MSE, and R2.

	Linear Regression	**Decision Tree**	**Random Forest**	**KNN**	**XGBOOST**
CC	0.04s	0.32s	27.08s	0.17s	32.86s
MSE	53748.440	25810.652	7260.332	0.05112	163.826
R^2	0.068	0.552	0.874	1	0.997
EV	0.068046	0.552464	0.874166	0.999	0.997164

Tableau 4 : Evaluation

III.4 Conclusion

This chapter is devoted to the realization of our proposed prediction system, which is divided into two parts. The first presents the hardware working environment, while the second details the results of the two sub-parts of our architecture, such as the machine learning part and the data visualization.

General conclusion

Artificial intelligence is nowadays an essential component for companies to exist in a very intense competitive environment. This report presents the work carried out which consisted in setting up a price analysis system in the Sonder platform.

In fact, this work focused primarily on the problem of prediction. Our system comprises three main processes: the first is price analysis, using Machine Learning algorithms.

To obtain the best-performing model, we used the following experimental protocols: computational complexity, variance explained, MSE and R2 statistic.

As for the results obtained, KNN and Xgboost gave us the best performance in terms of variance explained 99.7% and 99.9%, respectively.

This analysis system helps travelers choose their destination without the intervention of travel agencies.