**Airbnb Data analysis-Pyspark Project**

Airbnb is an online marketplace that connects people who want to rent out their homes with people looking for accommodations in that locale. NYC is the most populous city in the United States, and one of the most popular tourism and business places globally.

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Nowadays, Airbnb became one of a kind service that is used by the whole world. Data analysts become a crucial factor for the company that provided millions of listings through Airbnb. These listings generate a lot of data that can be analyzed and used for security, business decisions, understanding of customers’ and providers’ behavior on the platform, implementing innovative additional services, guiding marketing initiatives, and much more.

**Introduction**

In this project, we delve into the insights hidden within the "AB\_NYC\_2019.csv" dataset, which contains information about Airbnb listings in New York City. Our objective is to uncover valuable patterns, relationships, and trends in the data.

**Load data set in pyspark notebook**

The dataset we use is “New York Airbnb Open Data” from Kaggle. Using Spark, we’ll load the CSV file.

File Name- AB\_NYC\_2019.csv



**Project Scope and Objective:**

Our project aims to explore and analyse the Airbnb dataset to gain a comprehensive understanding of the following:

* Price distribution and trends
* Room types and their popularity
* Geospatial insights about neighbourhoods
* Correlations between key variables

**Dataset Overview:**

The "AB\_NYC\_2019.csv" dataset is a comprehensive collection of information related to Airbnb listings in NYC. With columns such as:

- id: Unique listing identifier

- host\_id: Unique host identifier

- neighbourhood\_group: NYC borough where the listing is located

- neighbourhood: Specific neighborhood within the borough

- room\_type: Type of room being listed

- price: Listing price per night

- minimum\_nights: Minimum number of nights required for booking

- number\_of\_reviews: Number of reviews for the listing

- reviews\_per\_month: Average number of reviews per month

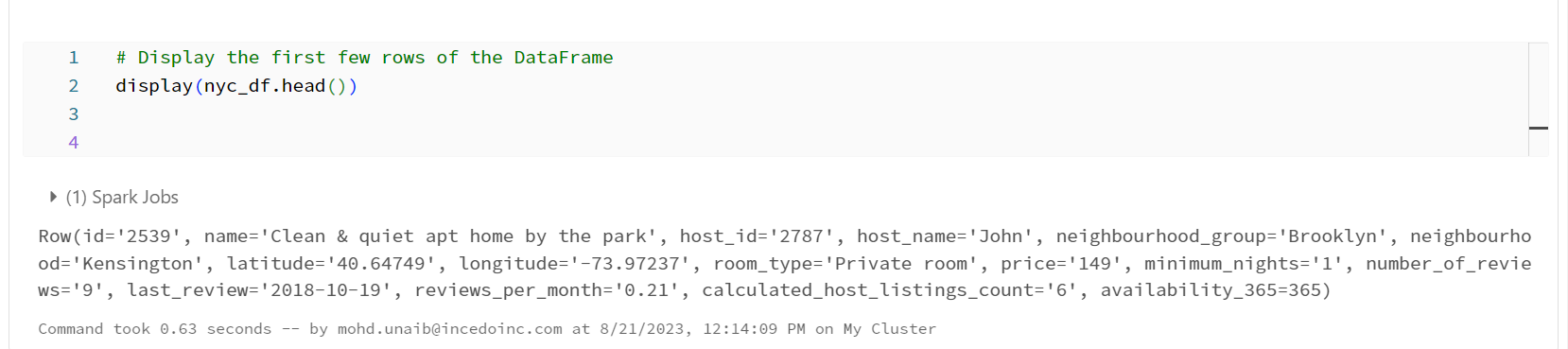
- calculated\_host\_listings\_count: Count of listings by the same host

- availability\_365: Number of days available for booking in a year



**Data Exploration Steps:**

1. Summary statistics to understand data distribution



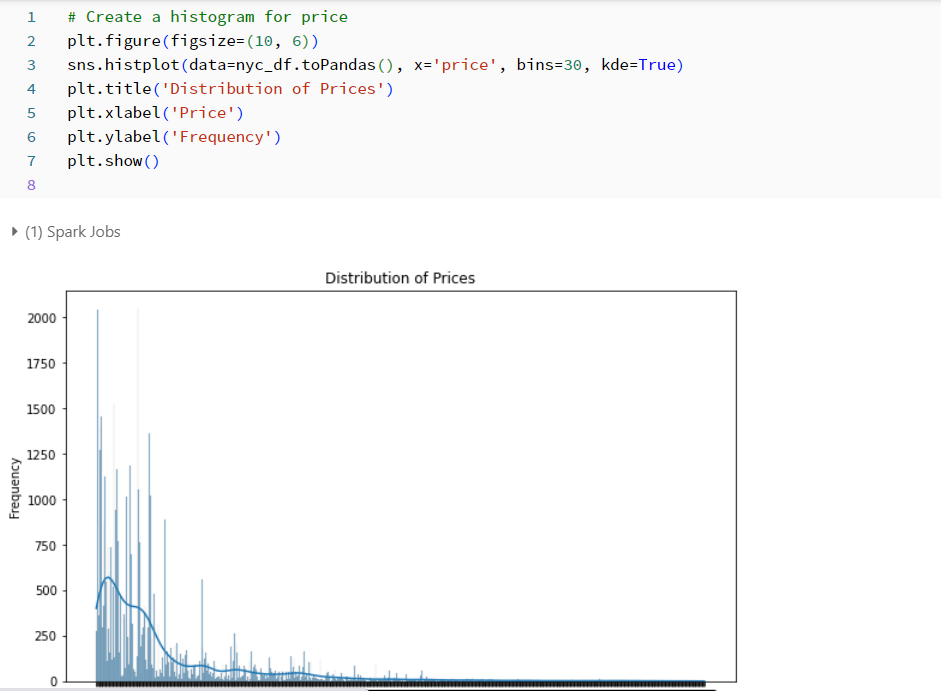


1. Handling missing values and their impact

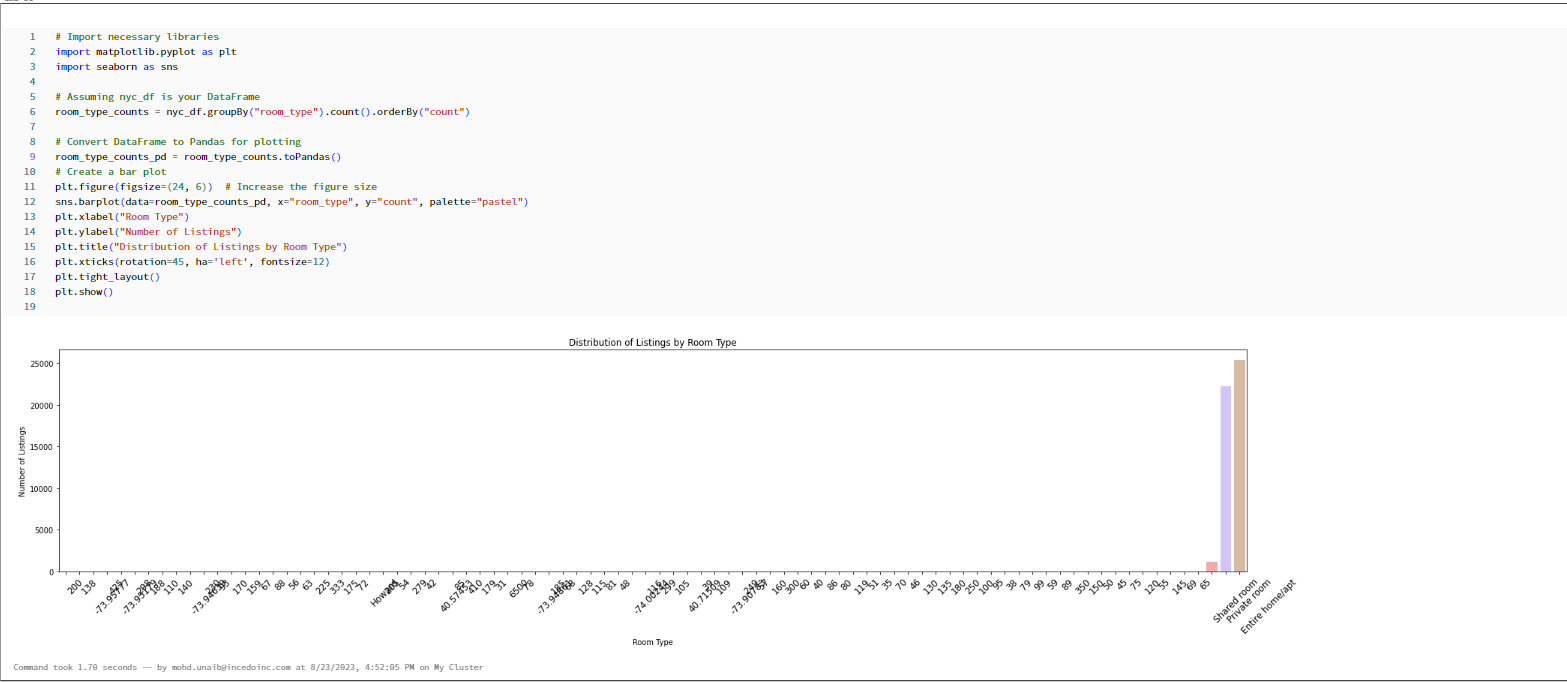




1. Visualizing price distribution through histograms



1. Analyzing categorical variables using bar plots



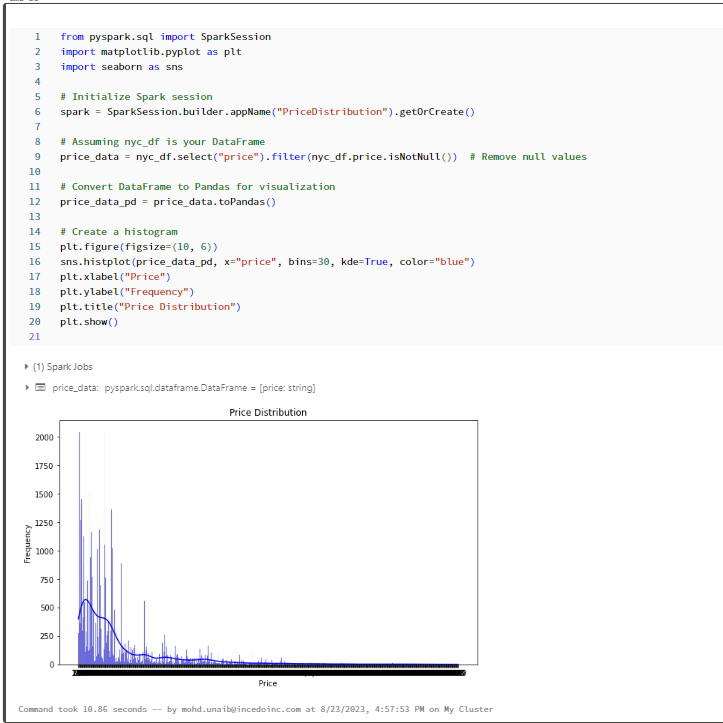
1. Exploring correlations between variables





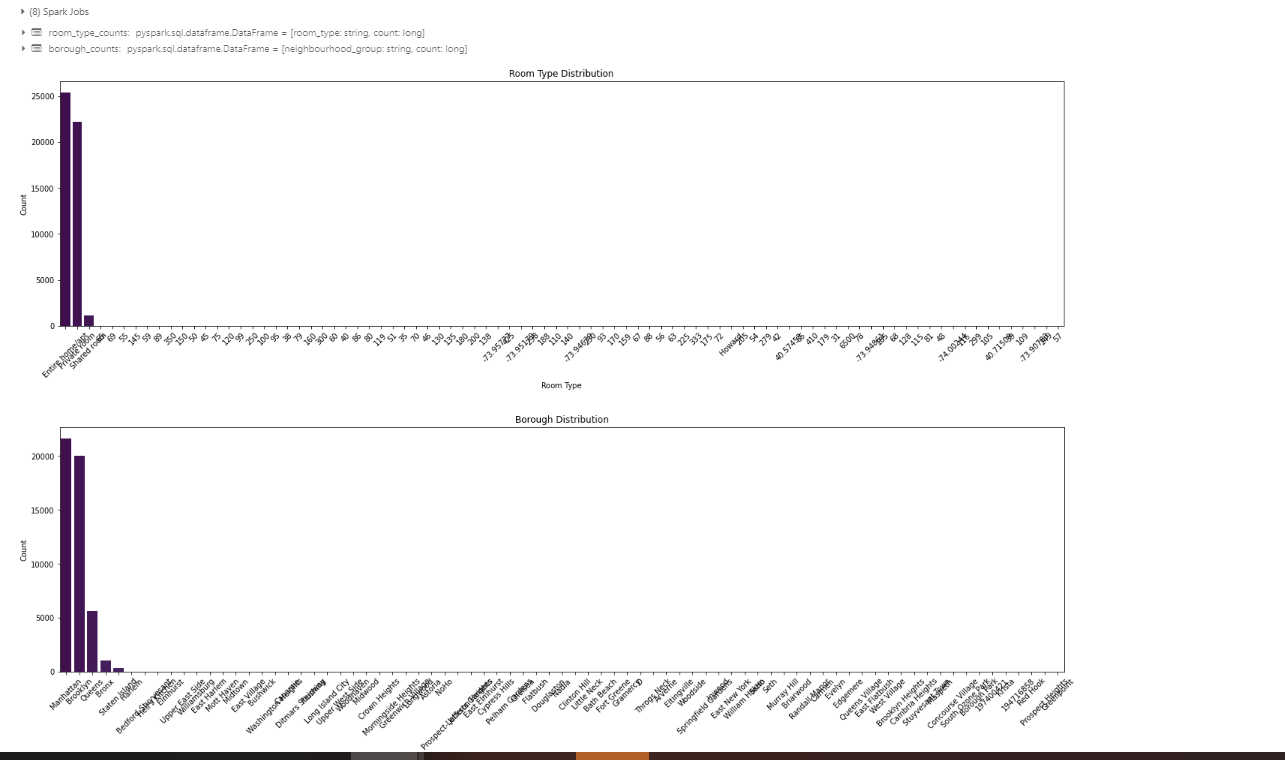
**Visualizations:**

1. Price Distribution: Histograms and density plots showcasing price spread

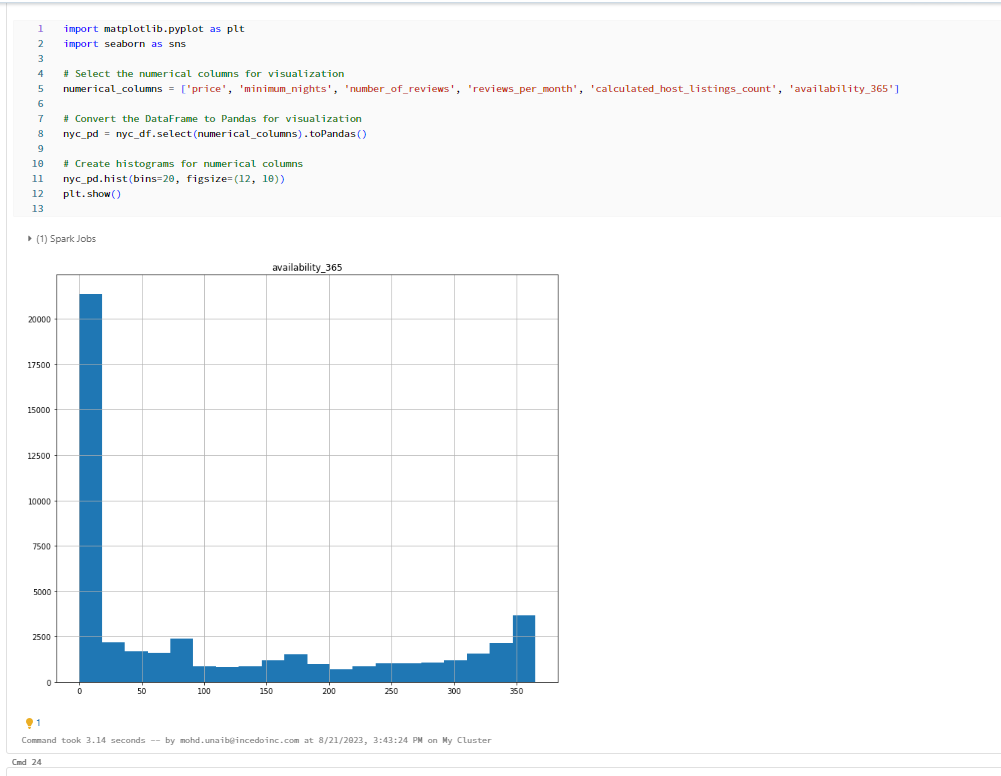


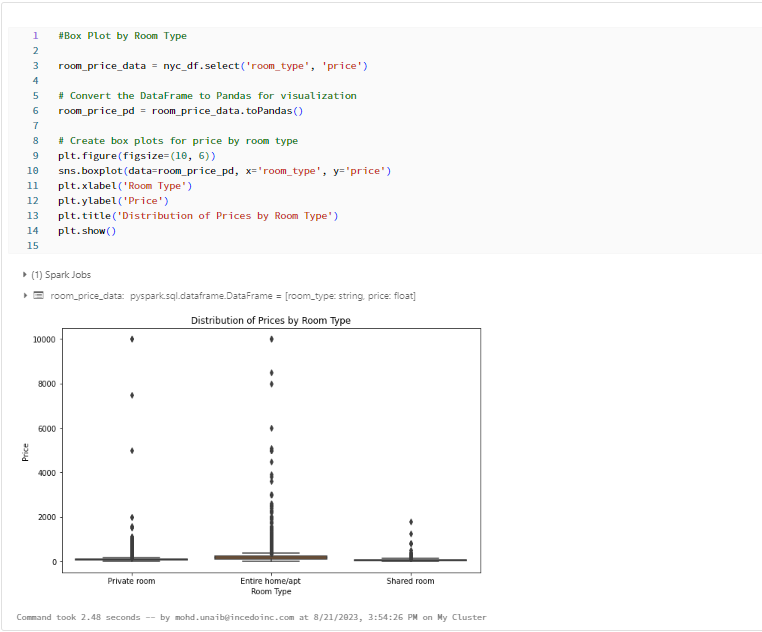
1. Categorical Analysis: Bar plots illustrating room types and borough distributions



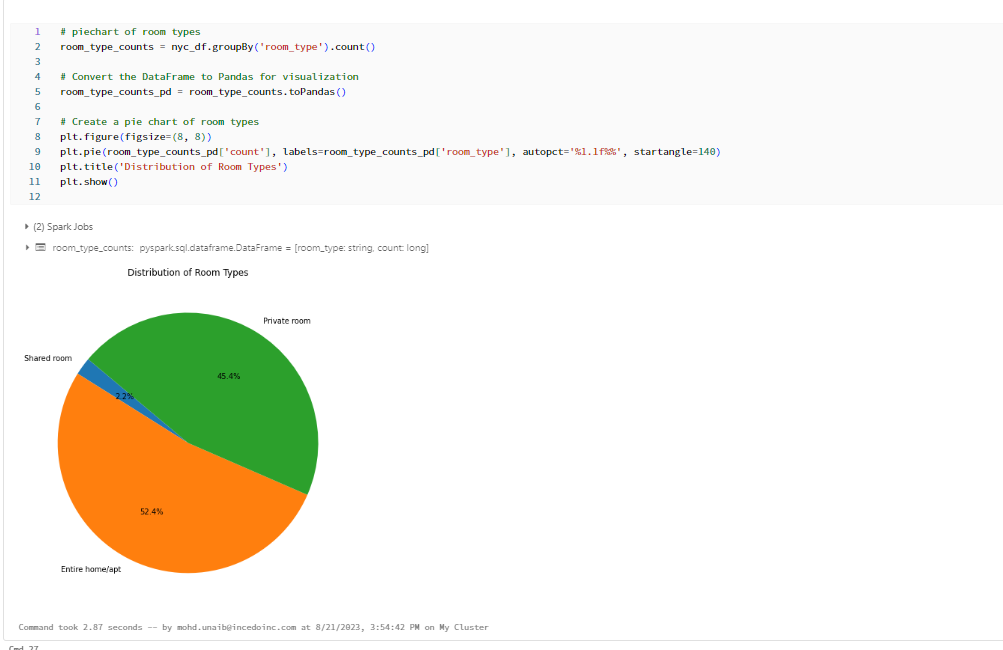


1. Relationships: Scatter plots revealing correlations between variables



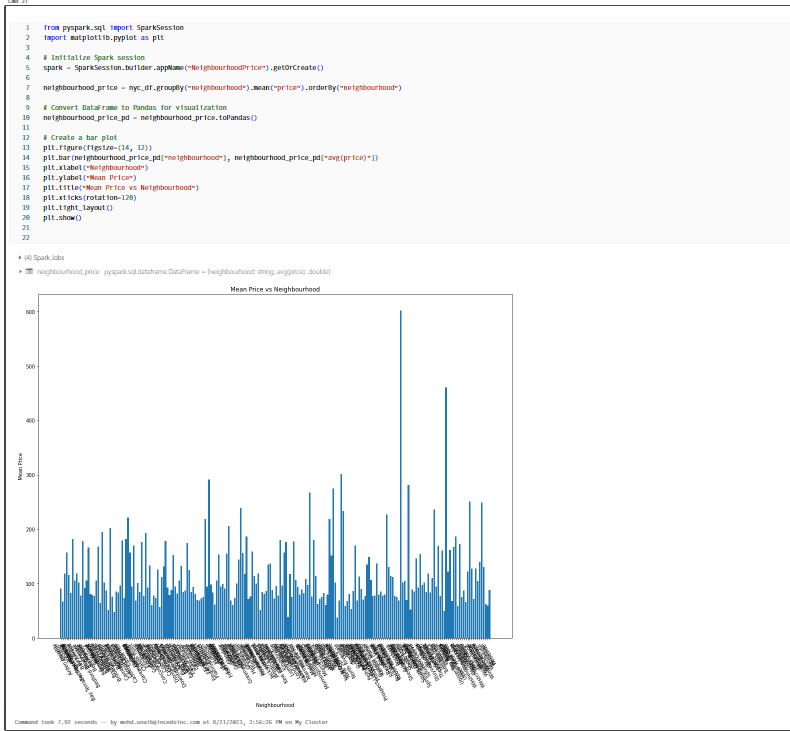






1. Mean Price by Neighbourhood:

We specifically analyzed the mean prices across different neighborhoods, which provided insights into the varying pricing trends.



**Conclusion:**

Through our data exploration, we have gained valuable insights that can inform decision-making and guide further analyses. The diverse factors affecting Airbnb listings provide a comprehensive picture of the NYC market.

**Future Steps:**

For future analysis, predictive modelling and time series analysis can further deepen our understanding of pricing trends and user behaviours.

MOHD UNAIB

Thank you.