



**Table of Contents**

[**Introduction** 2](#_heading=h.gjdgxs)

[**Project Goal** 2](#_heading=h.30j0zll)

[**Source Data** 3](#_heading=h.1fob9te)

[**Target Data** 3](#_heading=h.3znysh7)

[**Data Flow** 3](#_heading=h.2et92p0)

[**Database Used** 4](#_heading=h.tyjcwt)

[**Why Mysql Database?** 5](#_heading=h.3dy6vkm)

**# Unit 13| ETL Project**

**Topic: Futuristic Airbnb**

# **Introduction**

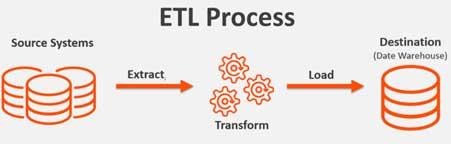
Airbnb is a home-sharing platform, which provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities. Guests can search for lodging using filters such as lodging type, dates, location, and price Guests have the ability to search for specific types of homes, such as bed and breakfasts, unique homes, and vacation homes before booking, users must provide personal and payment information. Some hosts also require a scan of government-issued identification before accepting a reservation Guests have the ability to chat with hosts through a secure messaging system Hosts provide prices and other details for their rental or event listings, such as the allowed number of guests, home type, rules, and amenities. The host, with recommendations from Airbnb, determines pricing. Hosts and guests have the ability to leave reviews about the experience.

# **Project Goal**

(**ETL**) is the general procedure of copying data from one or more sources into a destination system, which represents the data differently from the source(s) or in a different context than the source(s).

Our Project Aim is to perform the below activities as part of ETL phase.

* **Extract**is the process of *extracting data* from multiple and different types of sources.
* **Transform** isthe process of *converting the extracted data* from its previous form into the form it needs to be in so that it can be placed into a target database. Transformation occurs by using rules or lookup tables or by combining the data with other data.
* **Load** is the process of *writing the data* into the target database.



# **Source Data**

The Data Set Used for the Project:

* <http://insideairbnb.com/about.html>
* <https://www.kaggle.com/datasets>

The Data Files used for this project is as below.

* **listings.csv** 🡪 This file contains about the Airbnb Host, listing and Property Information’s. It also contains about the reviews scores & availability details of the property.
* **reviews.csv** 🡪 This file contains about the review information of the property.
* **calendar.csv** 🡪 This file contains about the Property availability and its price information for the future year.

# **Target Data**

The Final Cleaned Data flows in the below tables.

|  |  |
| --- | --- |
| **Table Name** | **Table Description** |
| **Futuristic\_Airbnb\_Hosts** | This table contains Host related Details. |
| **Futuristic\_Airbnb\_Listings\_Property** | This table contains the Listing and Property Details. |
| **Futuristic\_Airbnb\_Property\_Address** | This table contains the Property Address Details. |
| **Futuristic\_Airbnb\_Property\_Availability** | This table contains how many nights are available to be booked in the next 30, 60, 90, 365 days. |
| **Futuristic\_Airbnb\_Property\_Reviews** | This table contains the reviews of the Properties Details. |
| **Futuristic\_Airbnb\_Calendar** | This table contains the availability and price details for the future year. |

# **Data Flow**

























# **Database Used**

The Target Database Used in this project is **MySql**.

## **Why Mysql Database?**

# **Entity Relationship Diagram**

# **Source & Target Analysing Phase**

Before digging into the ETL Process, We first analysed the source thoroughly in order to split the source data into multiple target tables. The detailed analyzation is present in the below Table\_Identification sheet.



We have decided to split the source data to 6 tables like below.

**listings.csv**

Futuristic\_Airbnb\_Hosts  
Futuristic\_Airbnb\_Listings\_Property  
Futuristic\_Airbnb\_Property\_Address  
Futuristic\_Airbnb\_Property\_Availability  
Futuristic\_Airbnb\_Property\_Reviews

**reviews.csb**

Futuristic\_Airbnb\_Property\_Reviews

**calendar.csv**

Futuristic\_Airbnb\_Calendar

# **ETL Process**

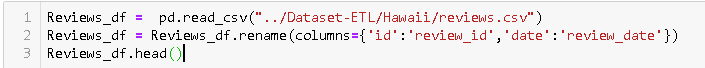
## **Extract**

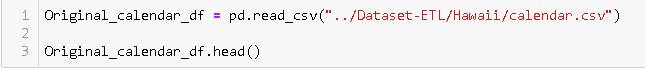
The below three CSV files are retrieved from Kaggle and insideairbnb websites.

* Listings.csv
* Reviews.csv
* Calendar.csv

Data in the above three files are extracted using pandas and stored in their respective data frames.







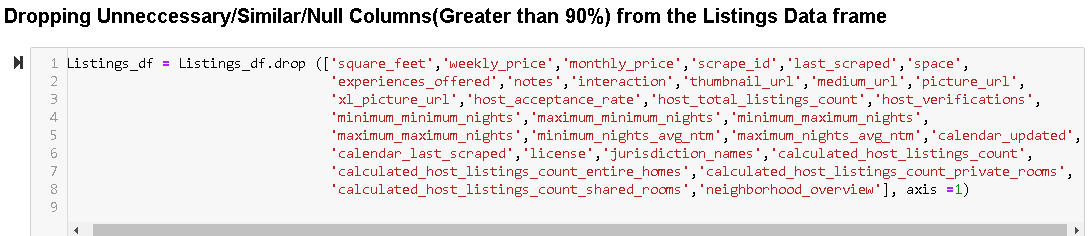
## **Transform**

Before Splitting the Data into 6 Dataframes, we first identified the unnecessary/duplicate/null value fields and dropped those fields from the main dataset.

**For Checking the Null Values in the Dataset :**



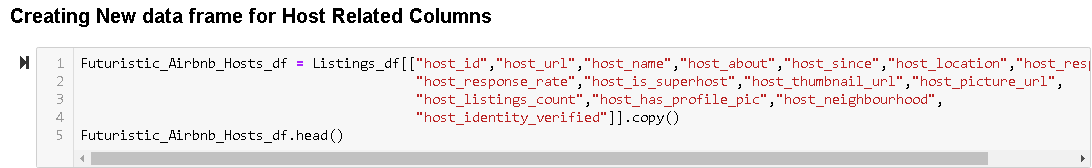
**For Dropping the Unnecessary/Similar/Null Fields in the Dataset :**



Since we are going to Split the above Source data into 6 tables, we have created 6 separate new data frames to store the respective fields like below.

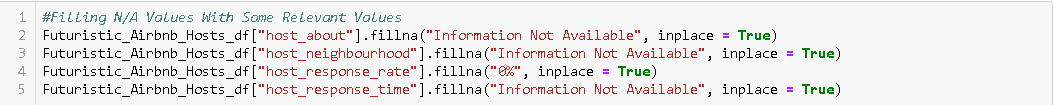
### **Futuristic\_Airbnb\_Hosts\_df**

***Step 1: Creating a New Data Frame for Holding the Host Related Columns***



***Step 2: Checking for the Null Values and Filling the N/A Values with Some Relevant Values***

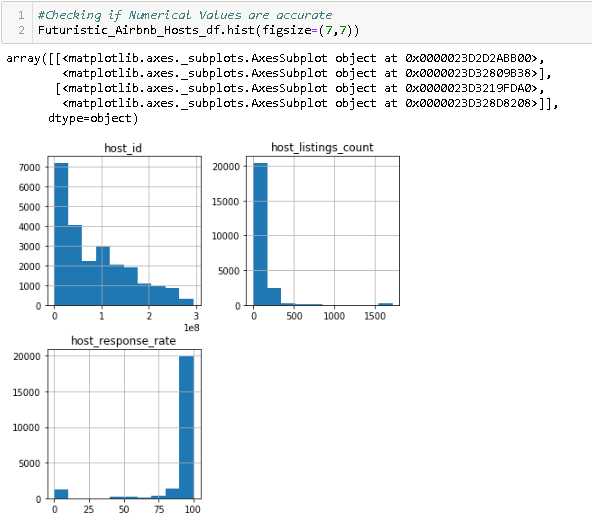


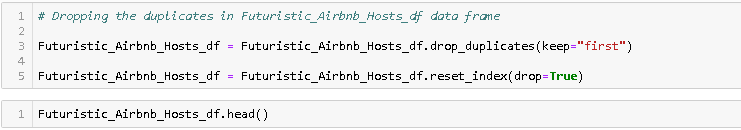


***Step 3: Data Clean up for the Respective Column have been done in the following ways:***

* ***Removing the % Symbol and converting to Numeric type***
* ***Converting the Boolean related columns to Boolean values as in mysql***
* ***Replacing the unwanted characters/words***
* ***Checking the accuracy of the numeric values by plotting histogram***
* ***Removing the duplicates***

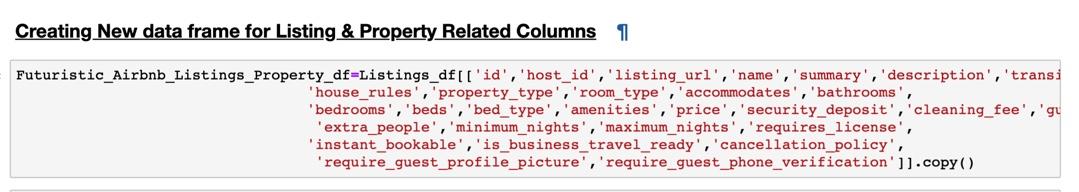




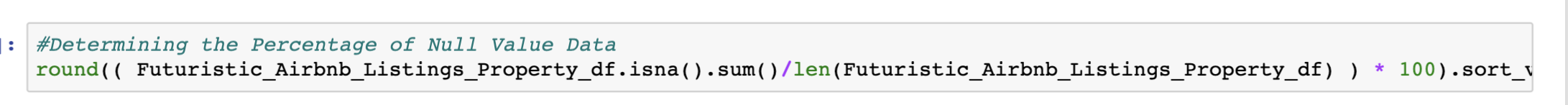


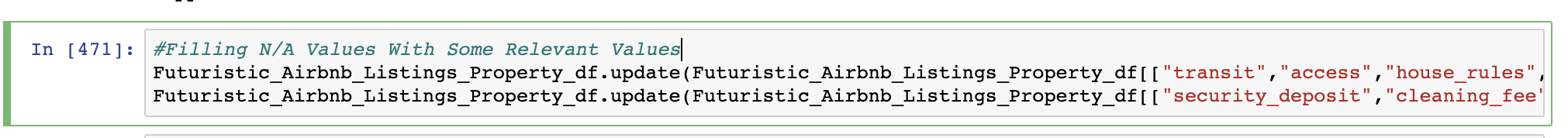
### **Futuristic\_Airbnb\_Listings\_Property\_df**

***Step 1: Creating a New Data Frame for Holding the Listing & Property Related Columns***

******

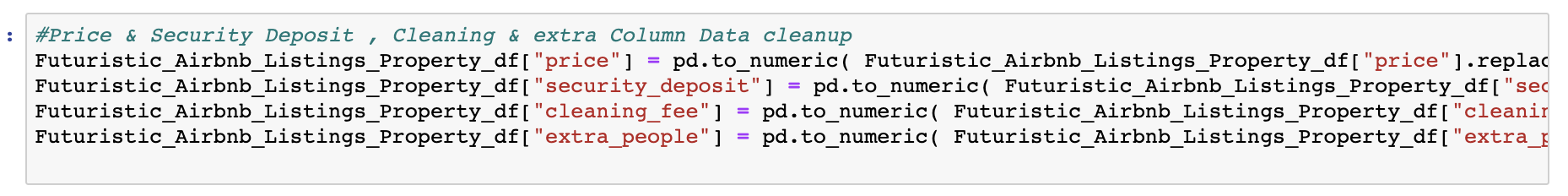
***Step 2: Checking for the Null Values and Filling the N/A Values With Some Relevant Values***

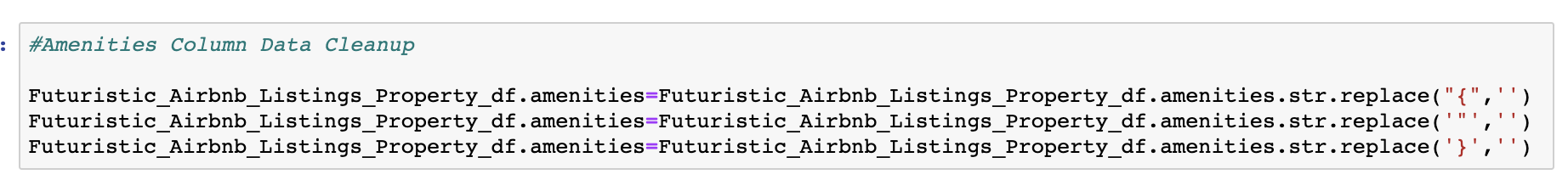
******

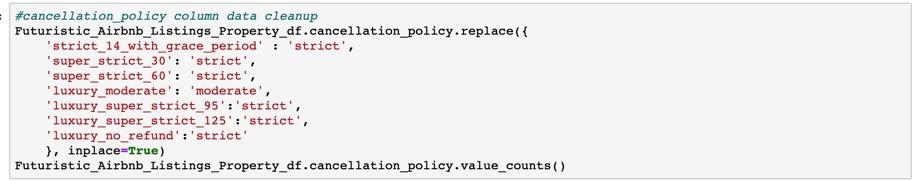
******

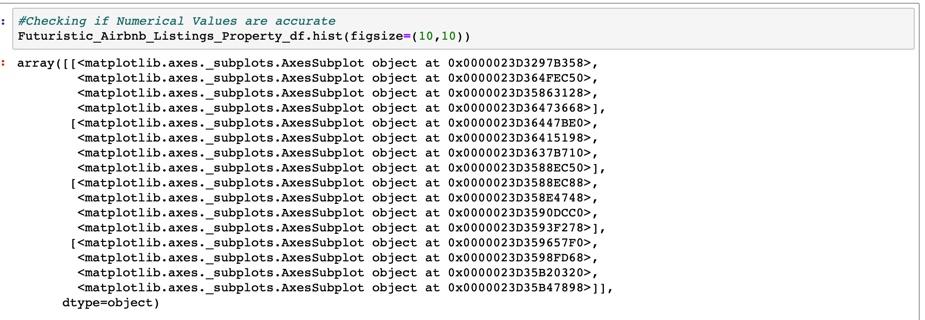
***Step 3: Data Clean up for the Respective Column have been done in the following ways:***

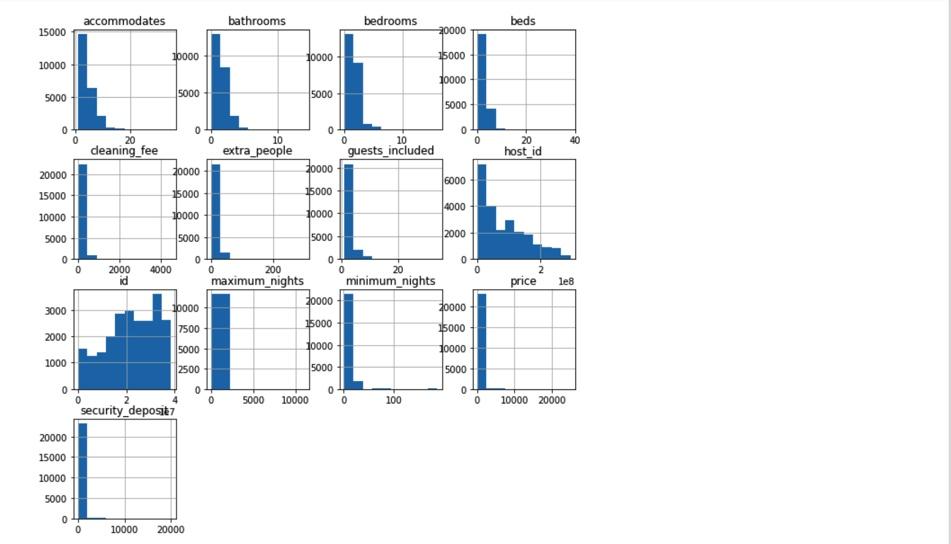
* ***Removing the % Symbol and converting to Numeric type***
* ***Converting the Boolean related columns to Boolean values as in mysql***
* ***Replacing the unwanted characters/words***
* ***Checking the accuracy of the numeric values by plotting histogram***
* ***Renaming column name***
* ***Removing the duplicates***

******

******

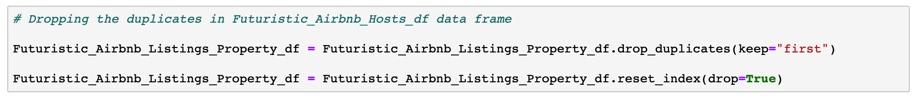
******

******

******

******

******

******

***Step 1: Creating a New Data Frame for Holding the Property Address Related Columns***

***Step 2: Checking for the Null Values and Filling the N/A Values With Some Relevant Values***

***Step 3: Data Clean up for the Respective Column have been done in the following ways:***

* ***Removing the % Symbol and converting to Numeric type***
* ***Converting the Boolean related columns to Boolean values as in mysql***
* ***Replacing the unwanted characters/words***
* ***Checking the accuracy of the numeric values by plotting histogram***
* ***Removing the duplicates***

### **Futuristic\_Airbnb\_Property\_Availability\_df**

***Step 1: Creating a New Data Frame for Holding the Property Availability Related Columns***

***Step 2: Checking for the Null Values and Filling the N/A Values With Some Relevant Values***

***Step 3: Data Clean up for the Respective Column have been done in the following ways:***

* ***Removing the % Symbol and converting to Numeric type***
* ***Converting the Boolean related columns to Boolean values as in mysql***
* ***Replacing the unwanted characters/words***
* ***Checking the accuracy of the numeric values by plotting histogram***
* ***Removing the duplicates***

### **Futuristic\_Airbnb\_Property\_Reviews\_df**

***Step 1: Creating a New Data Frame for Holding the Property Reviews Related Columns***

***Step 2: Checking for the Null Values and Filling the N/A Values With Some Relevant Values***

***Step 3: Data Clean up for the Respective Column have been done in the following ways:***

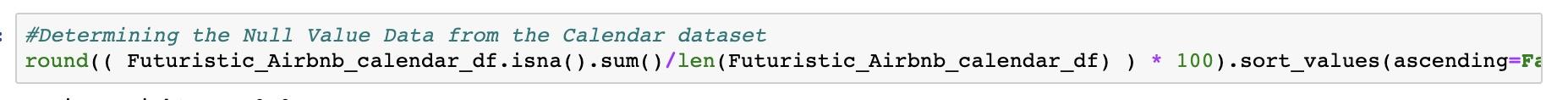
* ***Removing the % Symbol and converting to Numeric type***
* ***Converting the Boolean related columns to Boolean values as in mysql***
* ***Replacing the unwanted characters/words***
* ***Checking the accuracy of the numeric values by plotting histogram***
* ***Removing the duplicates***

### **Futuristic\_Airbnb\_Property\_Calendar\_df**

***Step 1: Creating a New Data Frame for Holding the Property Calendar Related Columns***

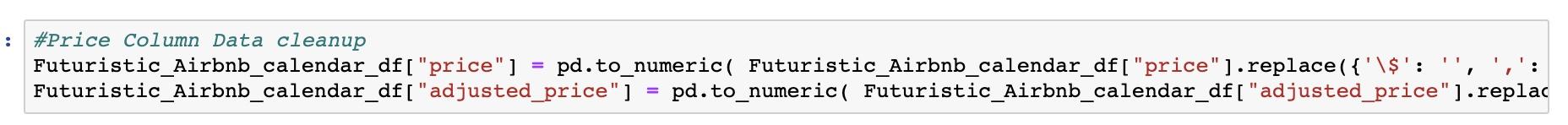
******

***Step 2: Checking for the Null Values and Filling the N/A Values With Some Relevant Values***

******

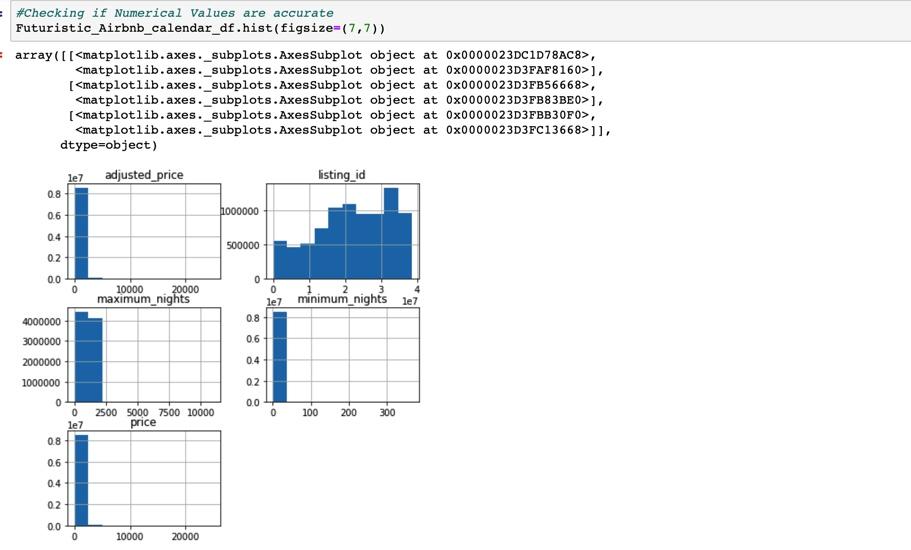
***Step 3: Data Clean up for the Respective Column have been done in the following ways:***

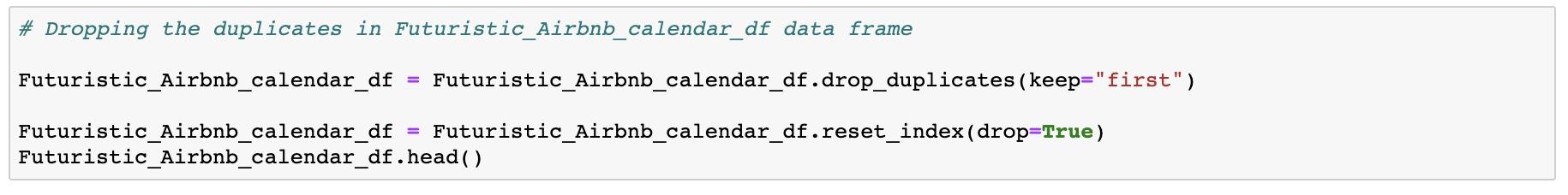
* ***Removing the % Symbol and converting to Numeric type***
* ***Converting the Boolean related columns to Boolean values as in mysql***
* ***Replacing the unwanted characters/words***
* ***Checking the accuracy of the numeric values by plotting histogram***
* ***Removing the duplicates***











## **Load**