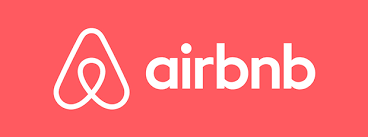
**METHODOLOGY DOCUMENT FOR STORY TELLING CASE STUDY: AIRBNB, NYC**

****

What is Airbnb?

Airbnb is a marketplace where travelers get to stay at the property listed on the site by the local hosts. Airbnb is a community built on sharing. It is based on an aggregator business model where the company collects the data about a specific good/service provider, make the providers their partners, and sell their services under its brand.

About The Dataset:

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values. Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other hand, Airbnb enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.

The data we are going to analyze is the data of Airbnb NYC (2019). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but not limited to, we will also try to explore some more insights.

Problem Background:

For the past few months, Airbnb has seen a major decline in revenue due to COVID situation. Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.So, analysis has been done on a dataset consisting of various Airbnb listings in New York.

The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset to increase the revenue such as –

Which type of hosts to acquire more and where?

The categorisation of customers based on their preferences.

* What are the neighbourhoods they need to target?
* What is the pricing ranges preferred by customers?

The various kinds of properties that exist w.r.t. customer preferences.

Adjustments in the existing properties to make it more customer-oriented.

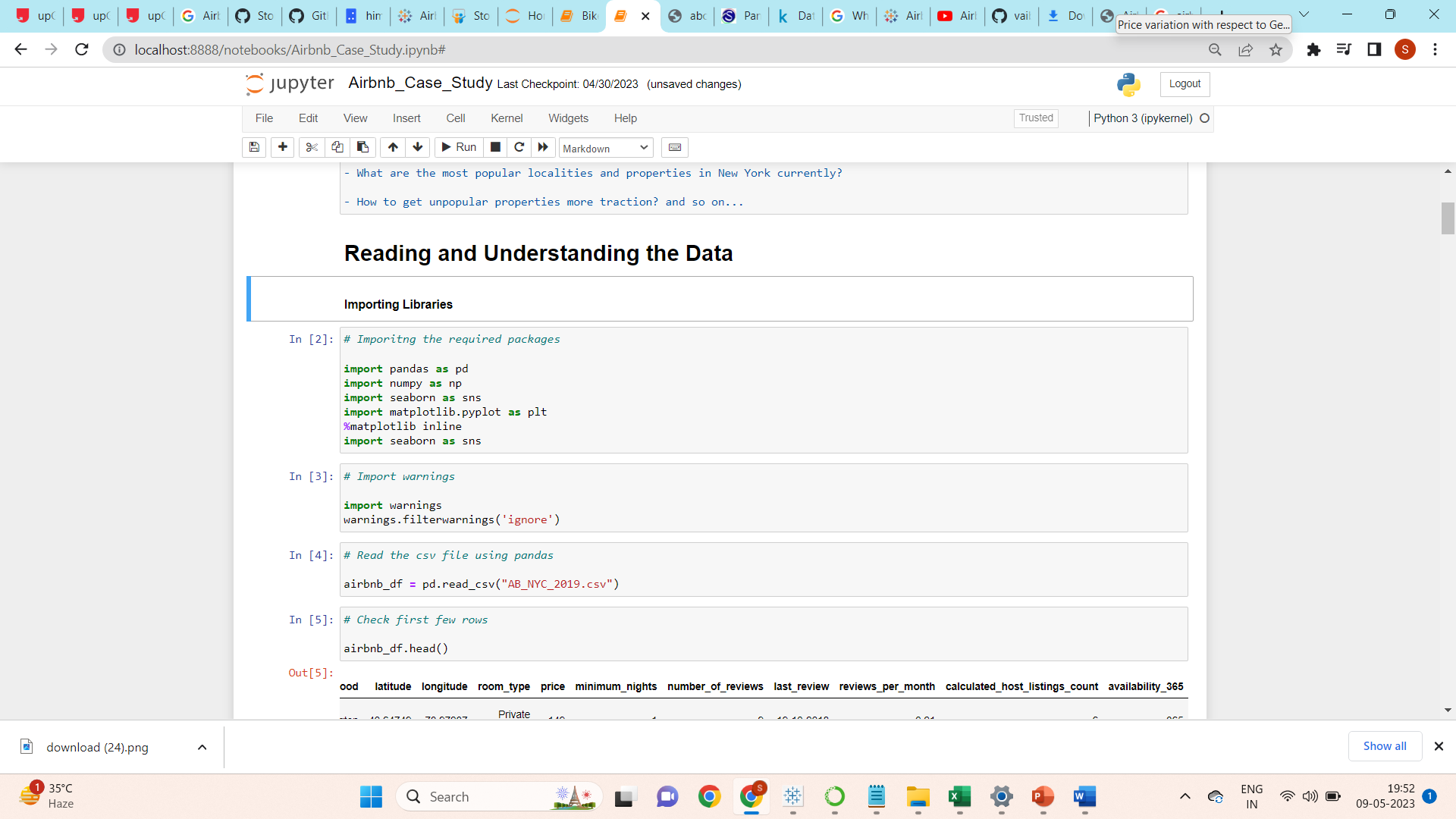
* What are the most popular localities and properties in New York currently?
* How to get unpopular properties more traction? and so on...

**Methodology Followed For EDA:**

The following steps are involved in the process of EDA:

* Acquire and loading data
* Understanding the variables
* Cleaning dataset
* Exploring and Visualizing Data
* Analyzing relationship between variables

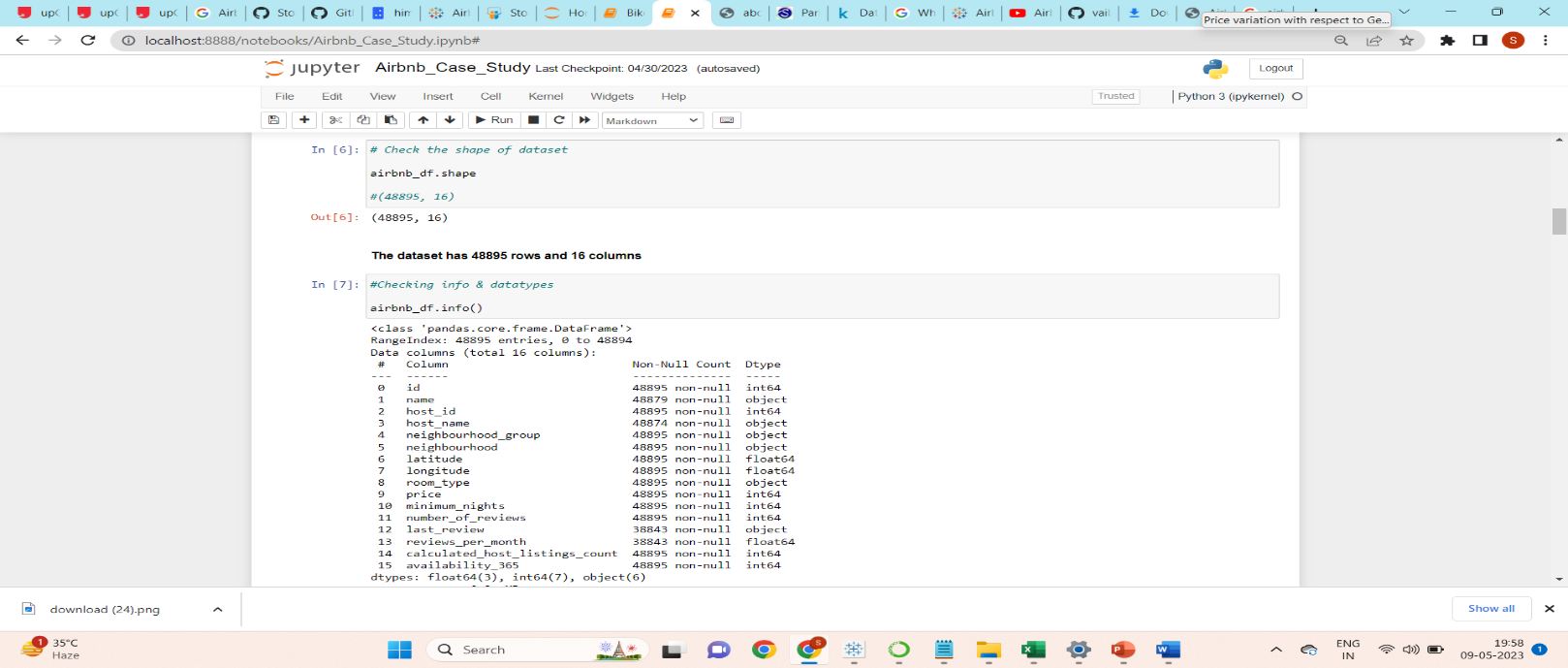
Reading and Understanding the Data: **Loading our data -** In this section we just loaded our dataset in python notebook and read the csv file using pandas library.



Understanding Data Variables:

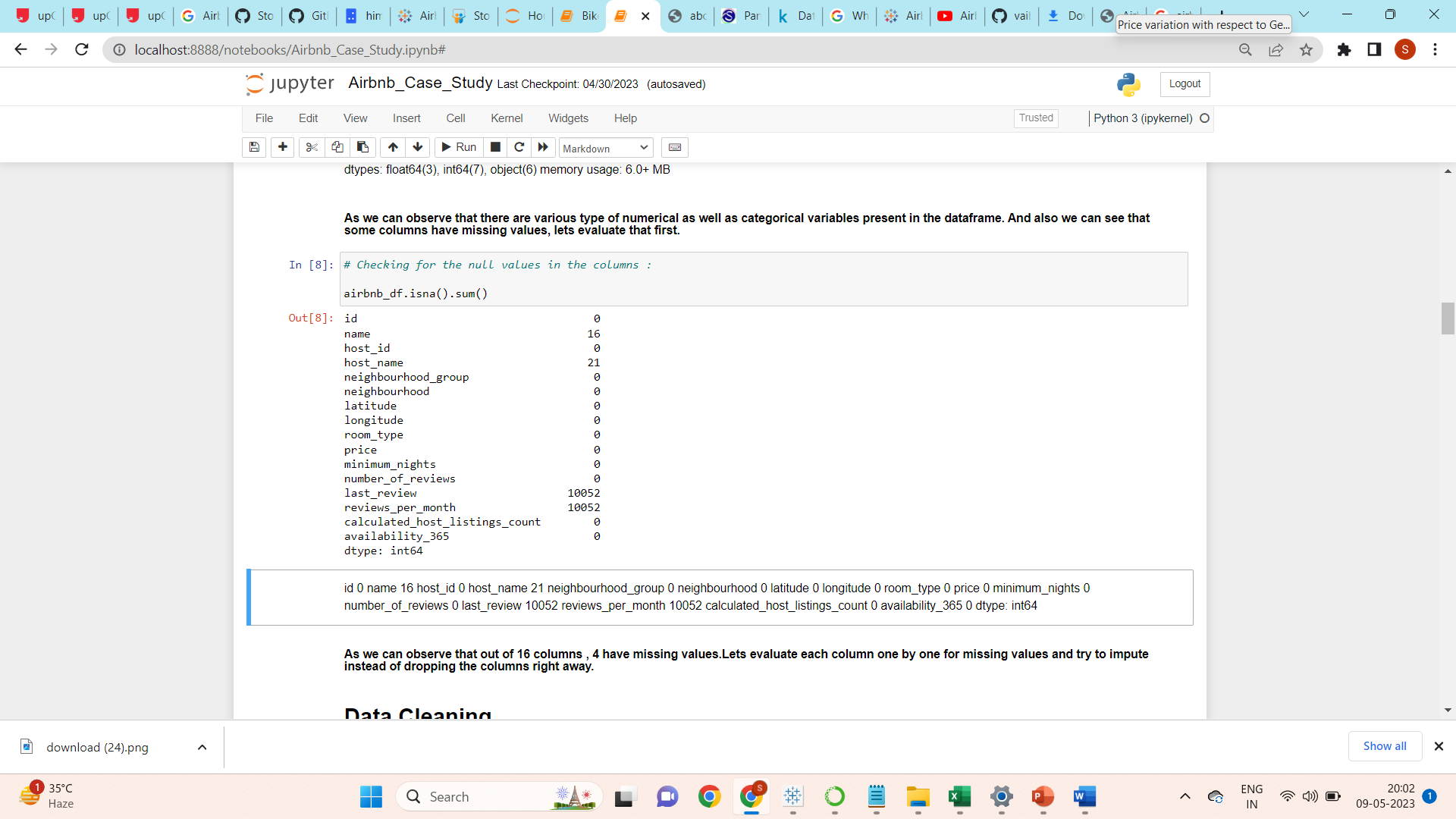
* **ID:**
  + It’s a unique id for House/apartment.
* **Name:**
  + Name of the listing House/apartment.
* **Host Id:**
  + Host Id is the government approved id for everyone who rent their properties on Airbnb.
* **Host Name:**
  + Host names are basically the name of the individual or organization Who own a room/apartment on Airbnb website.
* **Neighbourhood groups:**
  + Neighbourhood groups are the cluster of neighborhoods in the area.
  + There are about 5 boroughs in the state.
* **Neighbourhood:**
  + When searching for accommodations in a city, guests can filter by neighbourhood attributes and explore layers of professional-quality content, including neighbourhood maps, custom local photography and localized editorial, details on public transportation and parking, and tips from Airbnb’s host community.
* **Latitude:**
  + Latitude is the measurement of distance north or south of the Equator.
* **Longitude:**
  + Longitude is the measurement east or west of the prime meridian.
* **Room type:**
  + Airbnb has 3 categories for types of space:
    - **Entire House/Apartment**
    - **Private room**
    - **Shared room**
* **Price ($):**
  + The total price ($) of Airbnb reservation is based on the rate set by the Host, plus fee or costs determined by either the Host or Airbnb.
* **Minimum\_nights:**
  + Minimum night is criteria for booking that guest have to pay for book that House/room or apartment**.**
* **Number\_of\_reviews:**
  + Number of reviews of each host submitted by guest.
* **Last\_review:**
  + Latest review submitted by guest as feedback.
* **Reviews per month:**
  + Number of review Host get per month**.**
* **calculated\_host\_listings\_count:**
  + Amount of listing per host.
* **Availability 365:**
  + It is an indicator of the total number of days the listing is available for during the year.

Checking Data types:



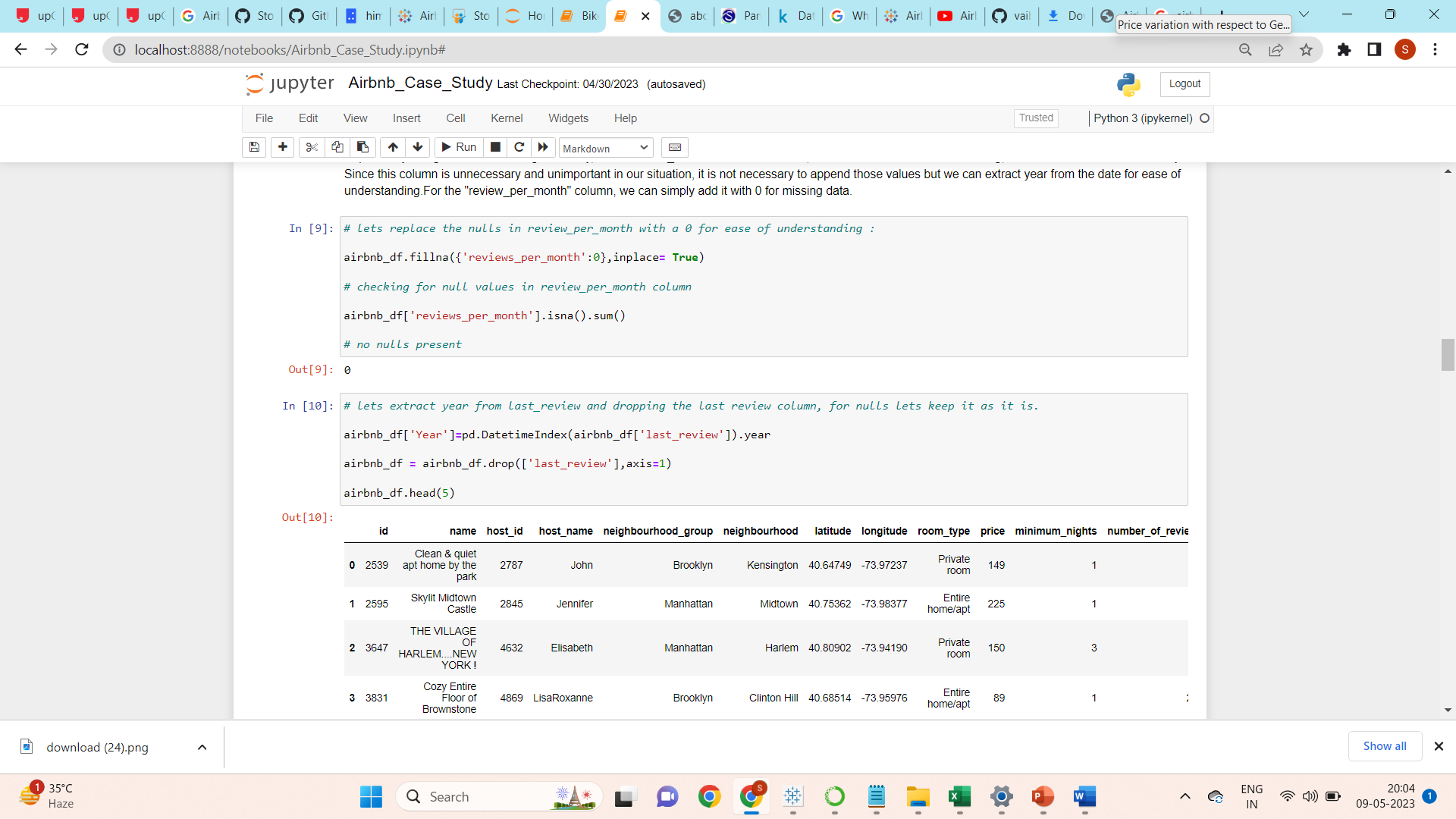
As we can observe that there is various type of numerical as well as categorical variables present in the dataframe. And also, we can see that some columns have missing values, lets evaluate that first.

Checking Null Values:

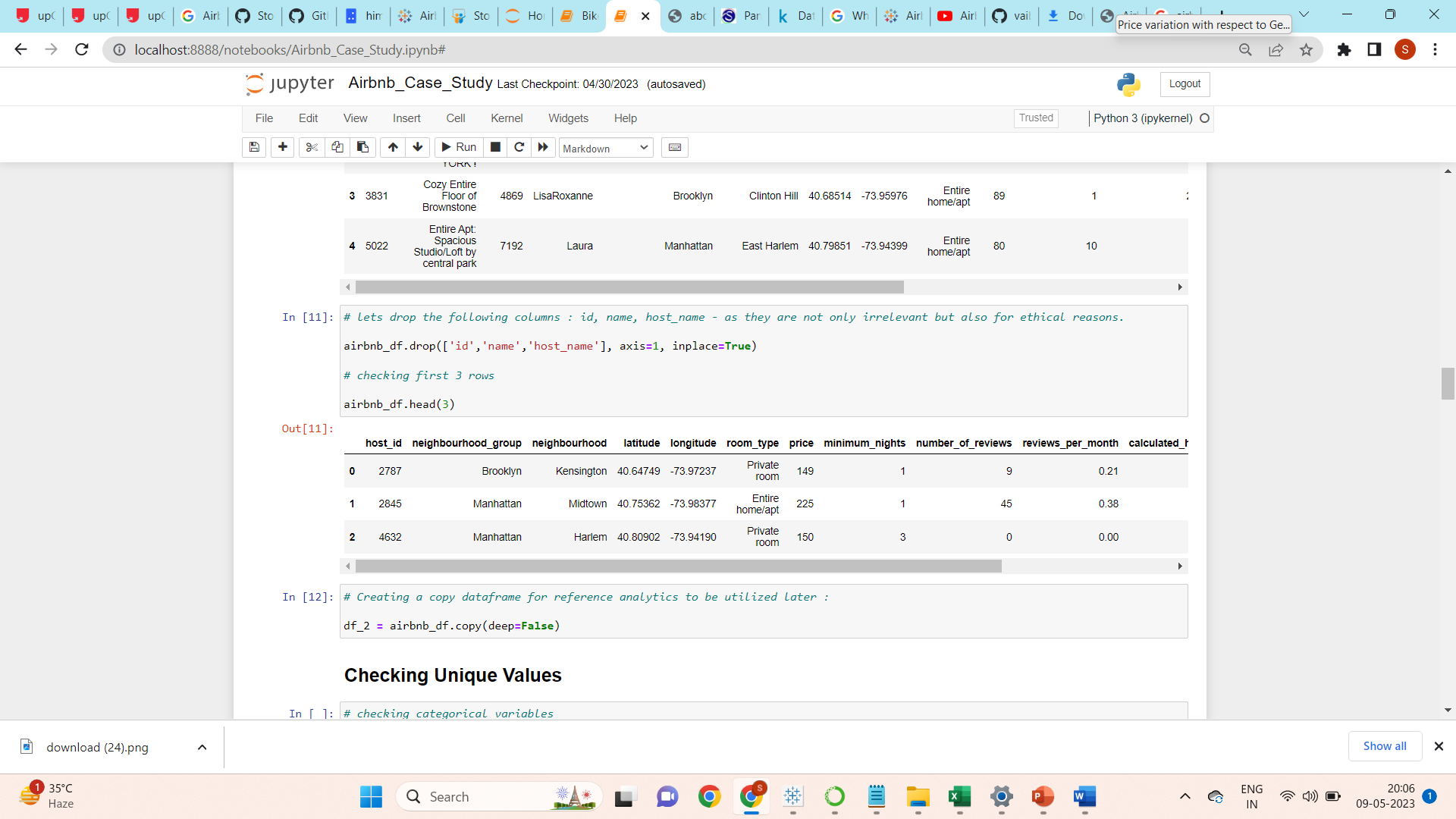


# Data Cleaning:

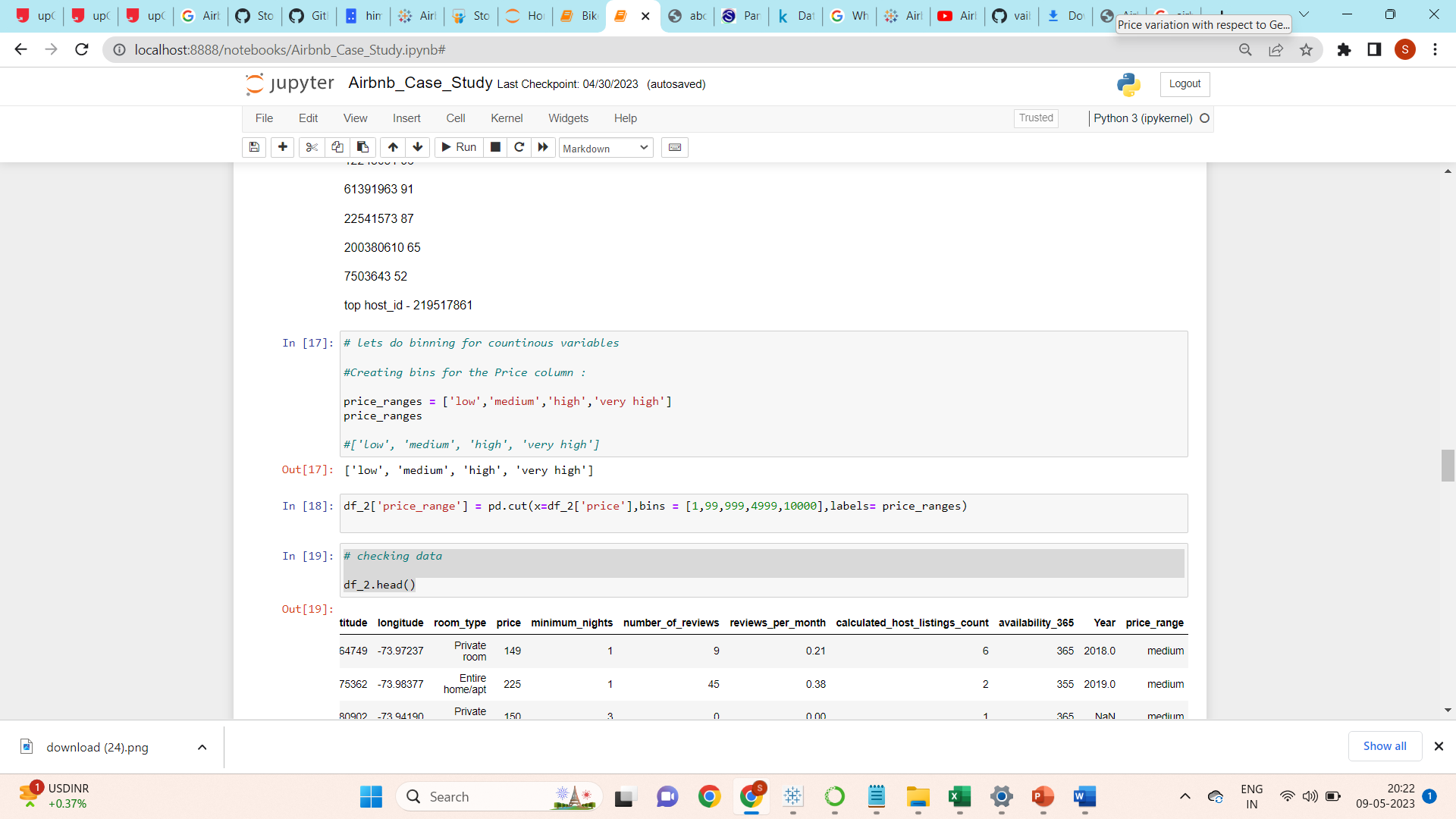
As we have observed that the missing data does not require a lot of extra consideration. Further observations can be made based on the characteristics of our dataset: the columns "name" and "host\_name" are unimportant and irrelevant to our data analysis, while the columns "last\_review" and "review\_per\_month" require very straightforward handling. To clarify, the term "last\_review" refers to the date; if there are no reviews for the listing, the data will be absent obviously. Since this column is unnecessary and unimportant in our situation, it is not necessary to append those values but we can extract year from the date for ease of understanding.For the "review\_per\_month" column, we can simply add it with 0 for missing data.



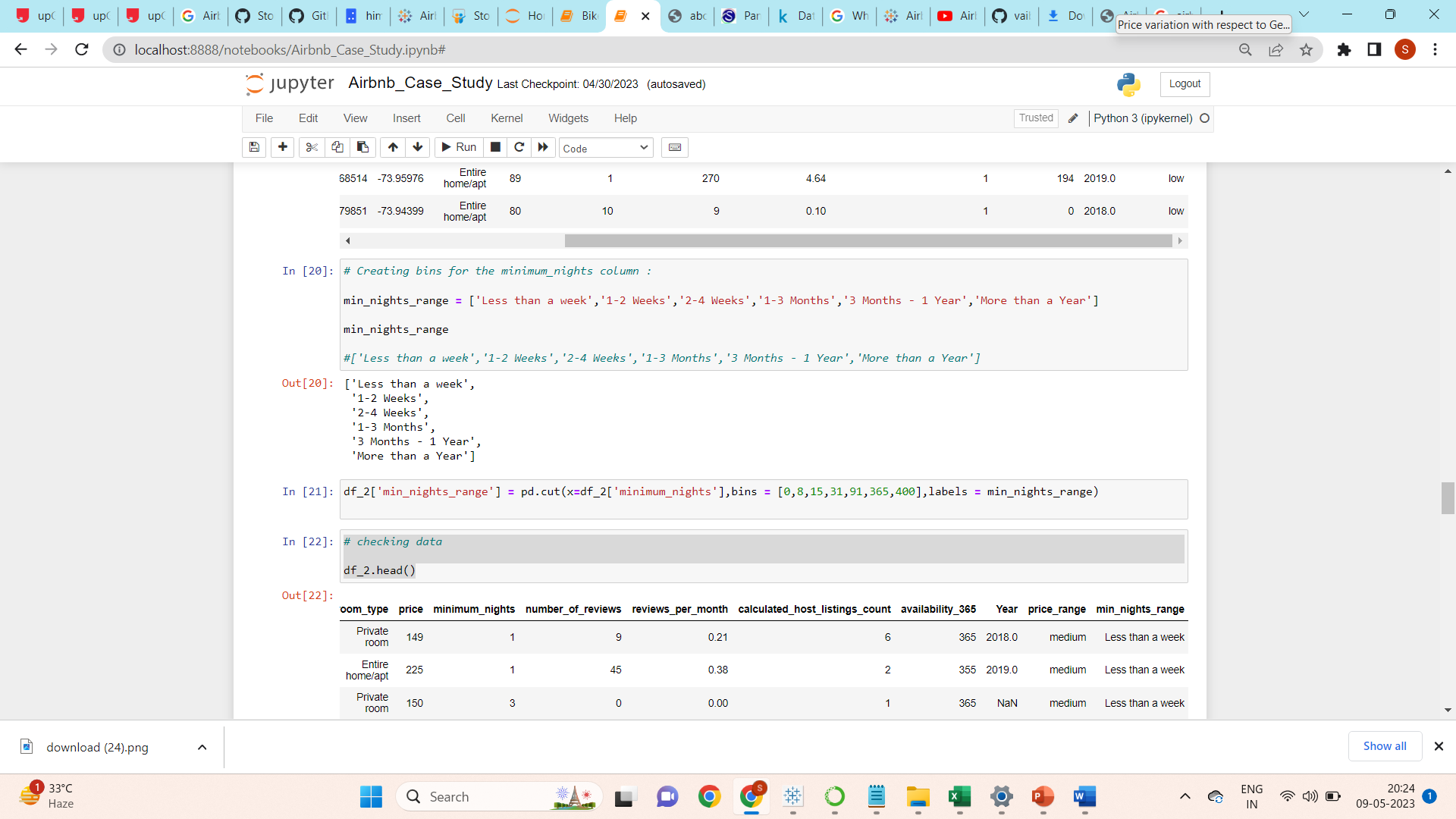
Dropping Irrelevant Columns:



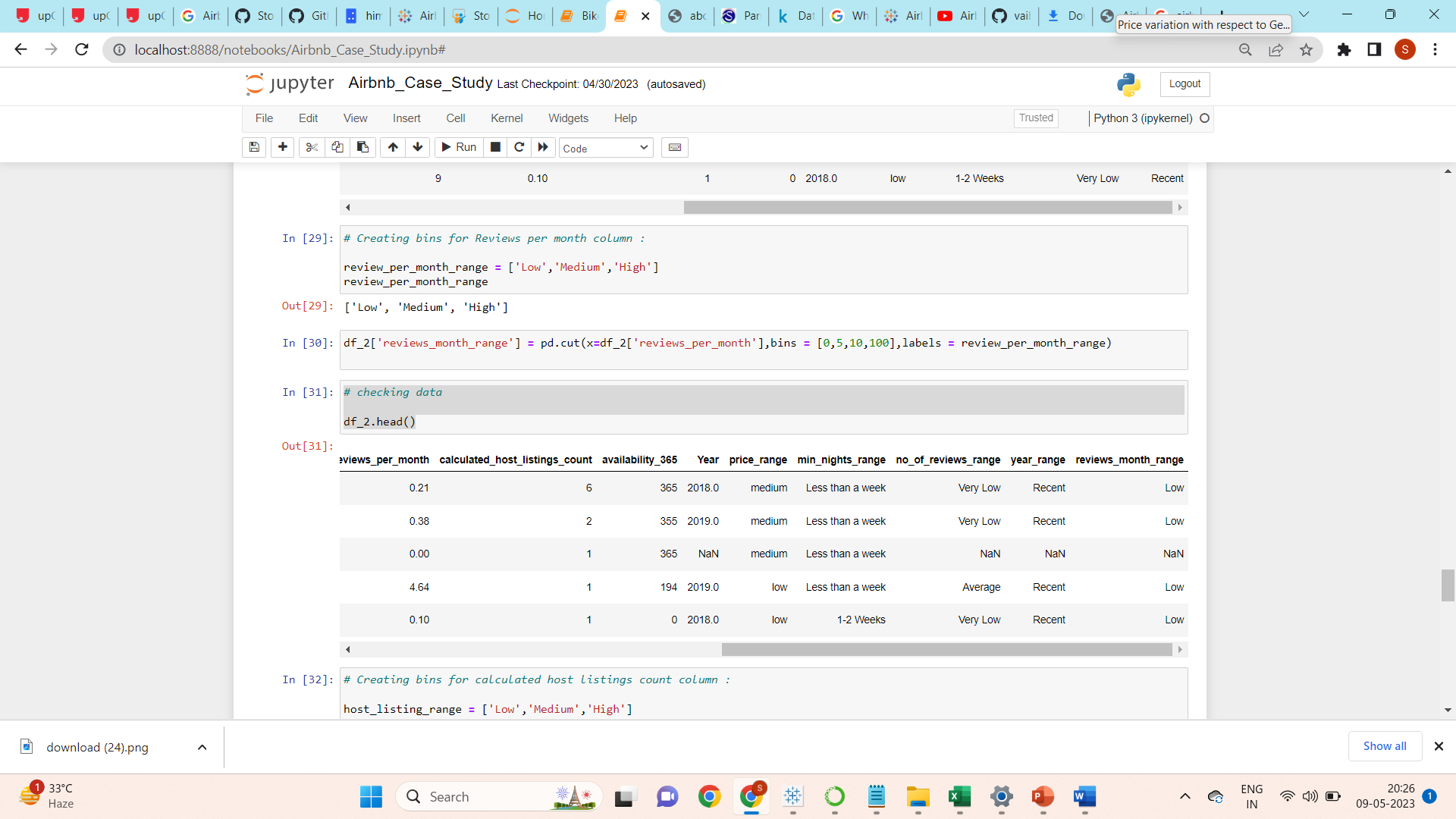
Created Bins for Continuous Variables:



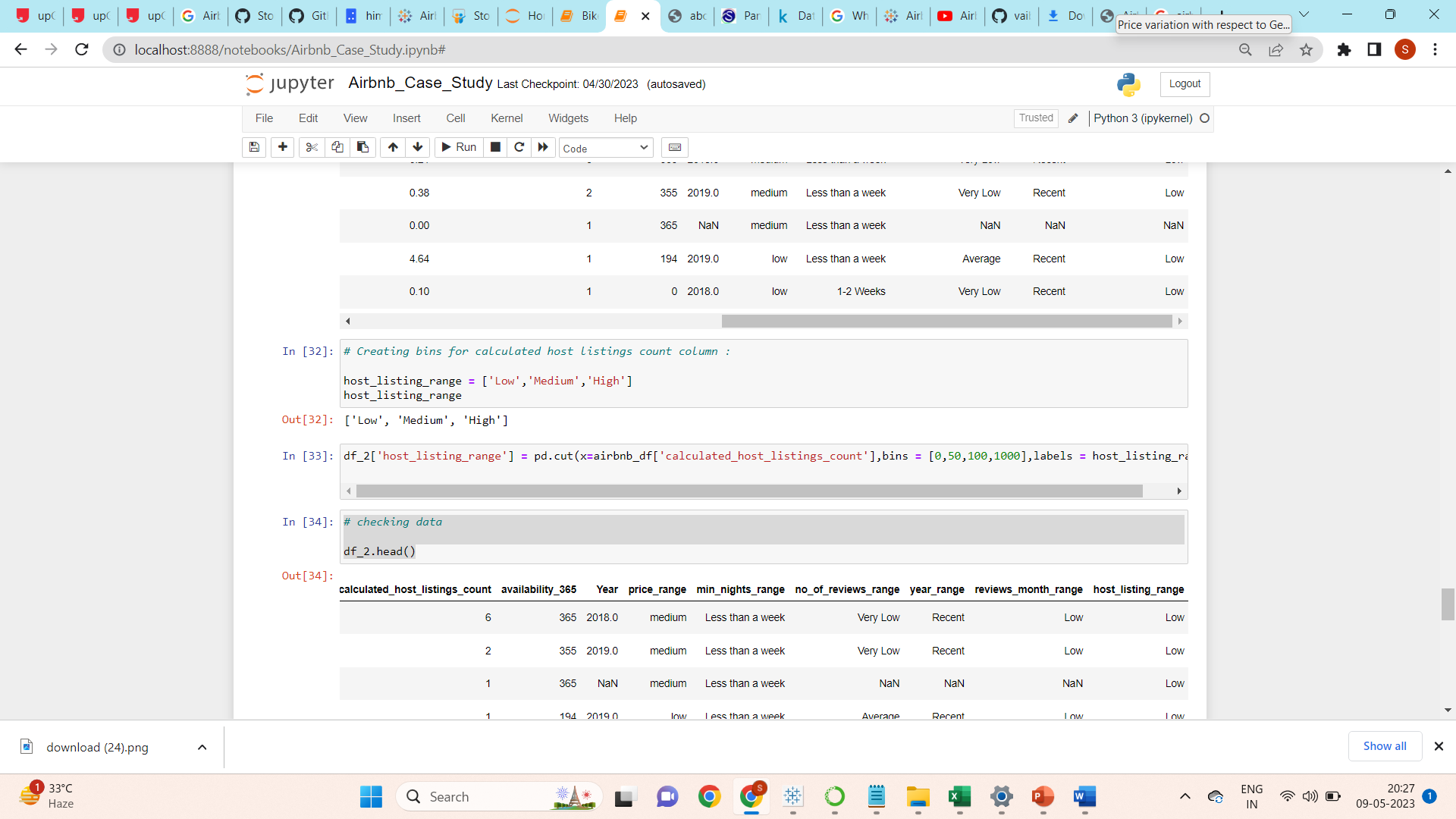
Bins created for price column as ['low', 'medium', 'high', 'very high']



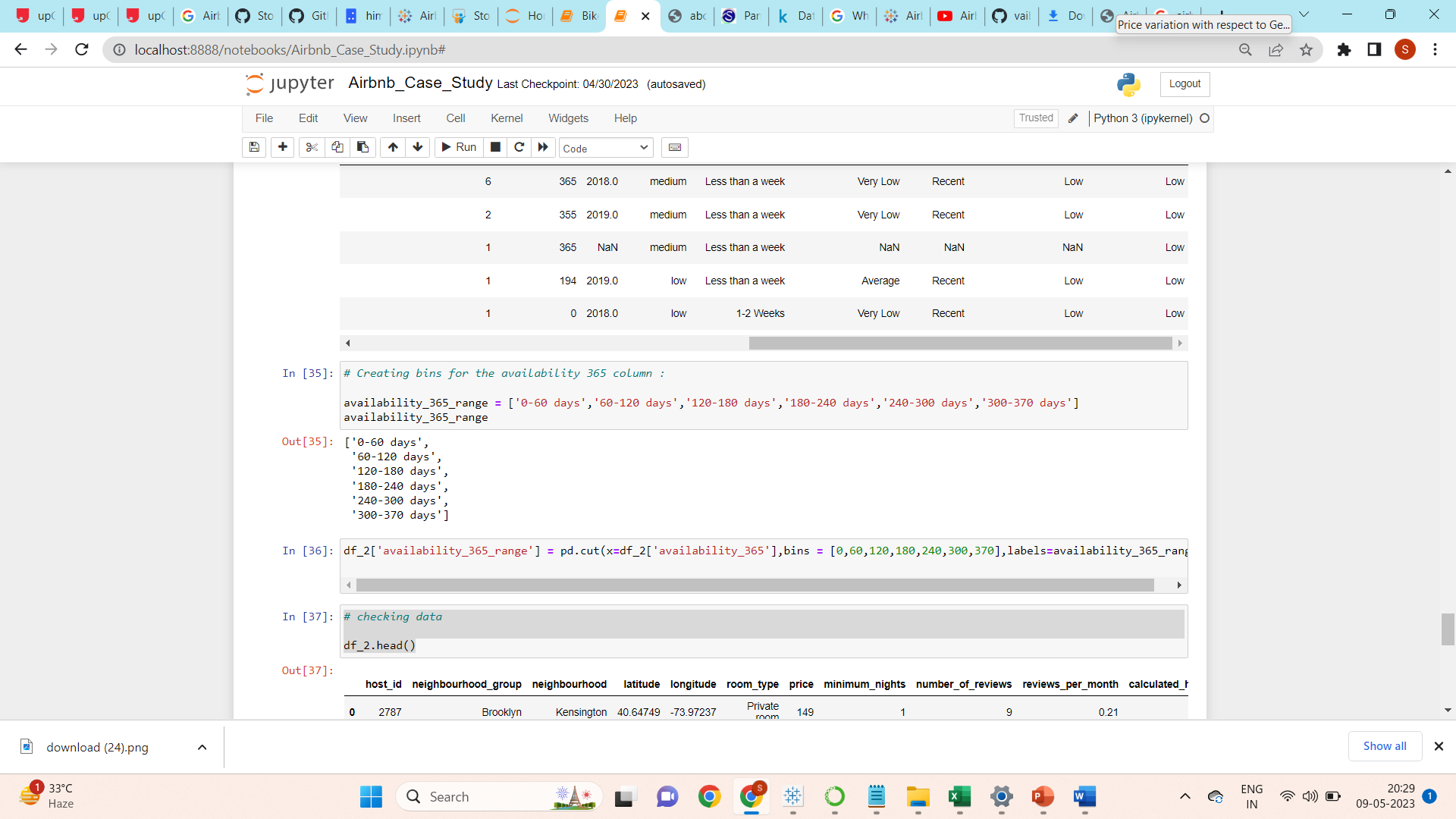
Bins created for Minimum nights column as ['Less than a week','1-2 Weeks','2-4 Weeks','1-3 Months','3 Months - 1 Year','More than a Year']



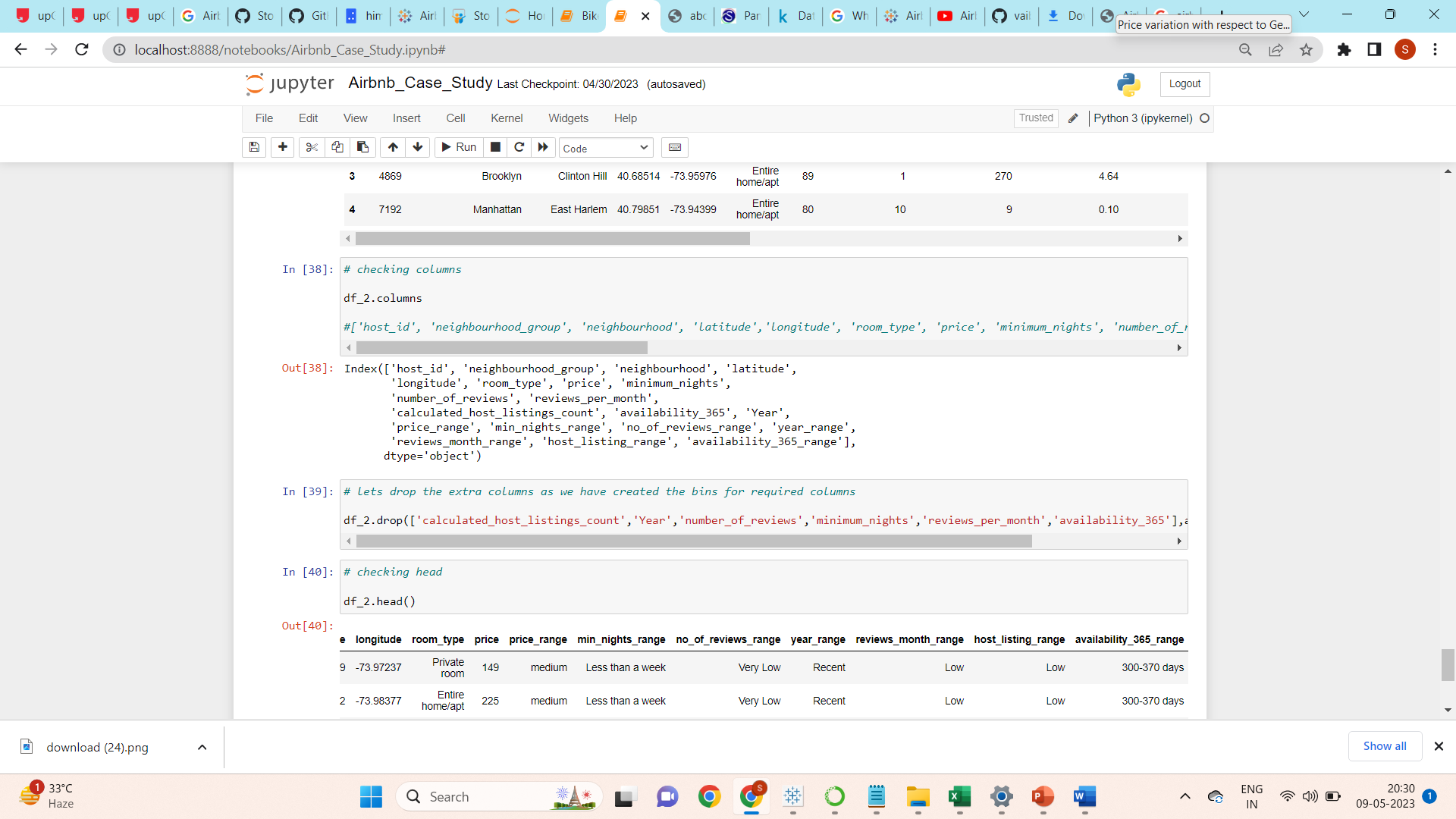
Bins created for reviews per month column as ['Low','Medium','High']



Bins created for calculated host listings column as ['Low','Medium','High']

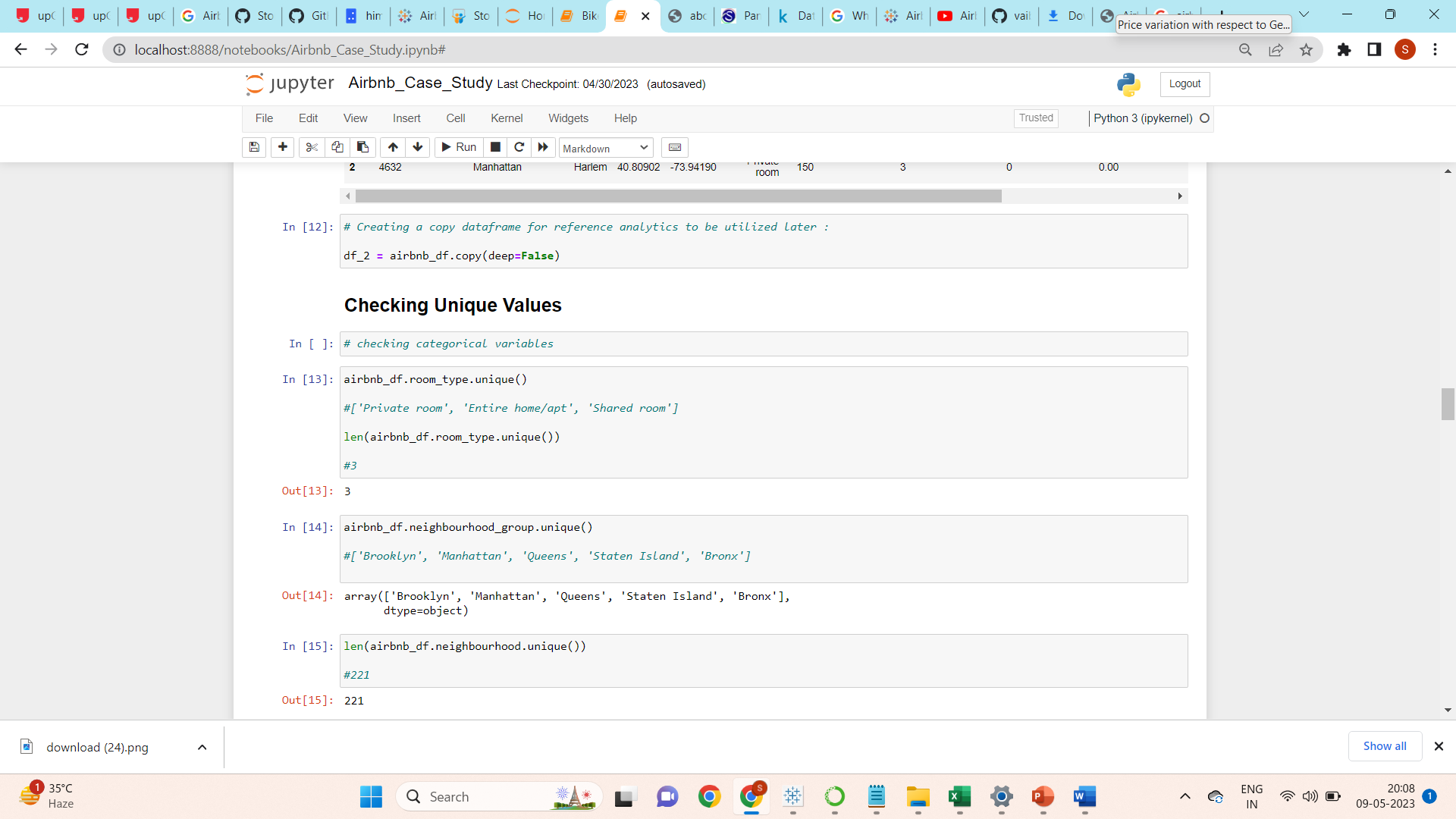


Bins created for availability column as ['0-60 days','60-120 days','120-180 days','180-240 days','240-300 days','300-370 days']

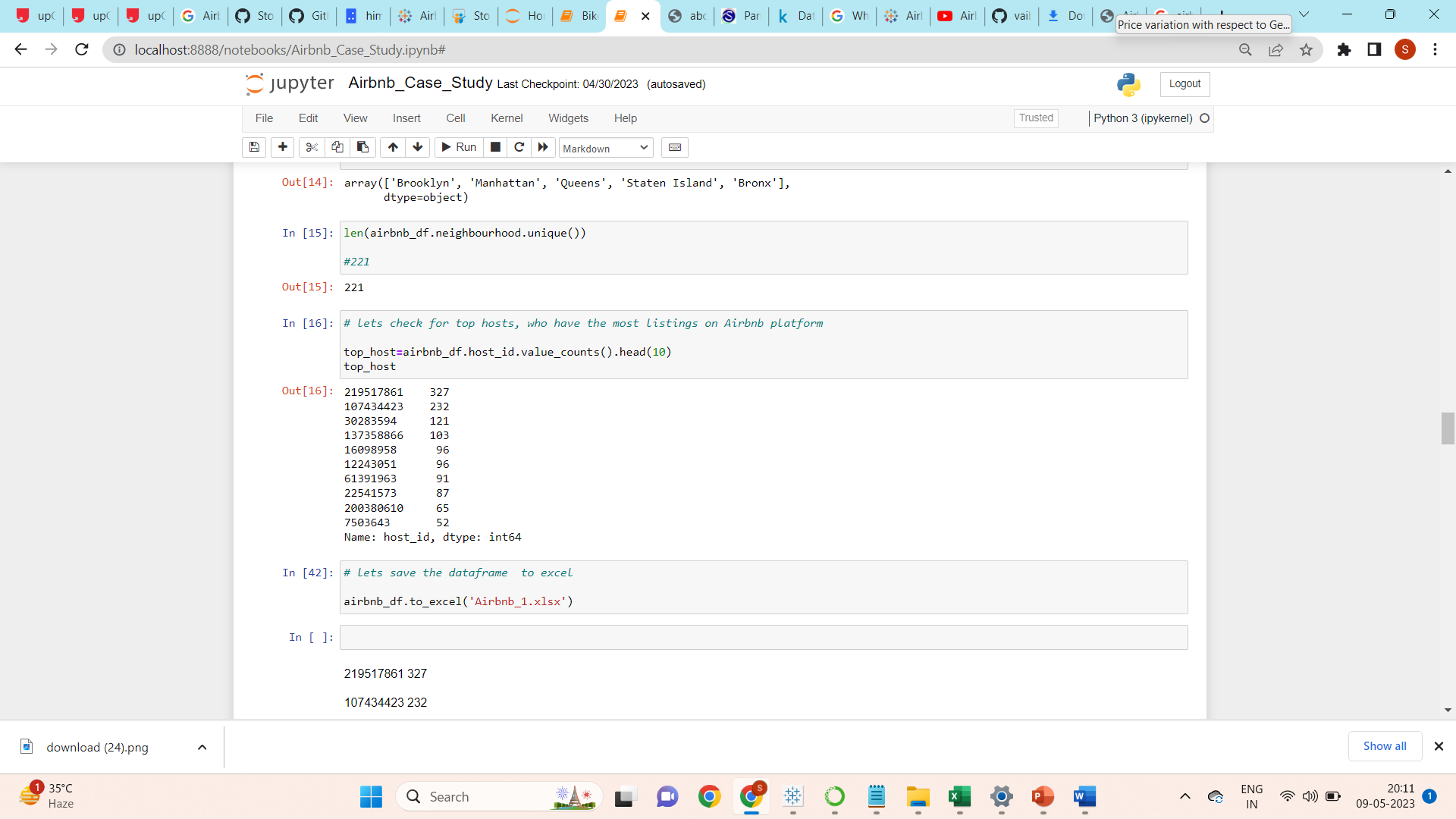


Checking columns & dropping the extra columns as we have created bins for required ones.

Checking Unique Values:



Checking Top 10 Hosts:



After understanding & doing the basic cleaning of data in python using pandas library, we have saved the dataset as Airbnb\_1.xlsx in excel file format. Further used this excel file to create visuals in Tableau.

The approach we have used in this project is defined in the given format.

1. **Loading our data: -** In this section we just loaded our dataset in python notebook and read the csv file using pandas library.
2. **Data Cleaning and Processing: -** In this section we have tried to check for the null values and for some of the columns we have replaced the null values with the appropriate values with reasonable assumptions.
3. **Analysis and visualization: -** In this section we have tried to explore all variables which can play an important role for the analysis using Tableau. In the next parts we have tried to explore the effect of one over the other. In the next part we tried to answers our hypothetical questions.
4. **Future scope of Further Analysis: -** There are many apartments having availability as **0**, which means they might stop their business, we can find the relation of neighbourhood with these apartments if we dig deeply, various micro trends could be unearthed, which we are not able to cover during this short duration efficiently. There are various columns which can play an important role in further analysis such as number of reviews and reviews per month finding its relation with other factors or other grouped factors can play an important role.

Files attached:

* PPT 1: Airbnb\_eda.pdf (ppt for technical leaders)
* PPT 2: Airbnb\_insights.pdf (ppt for business expertise)
* Python notebook: Airbnb\_Case\_Study.py
* Cleaned excel file: Airbnb\_1.xlsx
* Tableau workbook for visualizations: Airbnb\_visuals.twb
* Given Dataset: AB\_NYC\_2019.csv