# **Python Workshop - Data Cleaning and Exploration (3 hours)**

In this workshop you will practice your data exploration and analysis in Python. You will work in teams on this workshop.

1. Clean and Explore datasets in Python
2. Manipulate dataframes to answer questions
3. Use JupyterLab to explain your analysis

# **Part 1**

Read the dirty\_data.csv dataset into a dataframe in a new Jupyter notebook. Clean the data, paying attention to:

* Stray characters (you may need to use the escape \ for certain characters eg. ‘\\*’)
* Misspelled words
* Duplicate data
* Datatypes
* Null values

# **Part 2**

Explore the dataset. Use these questions to guide your exploration :

* How are the numeric values distributed (i.e. descriptive statistics)?
* What’s the average room price in each borough?
* What neighborhood has the most expensive rentals (max of mean)?
* What type of room costs the most? The least?
* Are higher ratings [correlated](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html#pandas.DataFrame.corr) with higher prices?
* Are hosts with more rentals more likely to receive higher ratings?
* What’s the ‘best’ rental in NYC, considering location, price, and rating? (this is a subjective judgment, but make sure you justify your choice)

# **Part 3**

Clean up your notebook (i.e. hide unnecessary cells) for presentation. Present your findings as a group. Add markdown explanations as needed.

We’ll convene 3:00 - 4:30 pm to discuss and share.

Teams assignments:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Room 1** | **Room 2** | **Room 3** | **Room 4** | **Room 5** |
| Shaday Brown | Chrissy Taylor | Angelica Vera | Sol-Marie Quintero | Kristian McCombs |
| Adam Shabana | Chhaya Penn | Ali Ashfaq | Brenda Jerez | Ess Guernah |
| Diana Ospina | Kosta Louvros | Victor Mantilla Colon | Chad Crossman | Daphney Oliveira |
| Marianna Beaute | Rosana Infante | Rosemary Espinal | Aaron Potts | Timothy Yip |
| Marcus Madison |  | Virna Brown |  |  |