# **Python Independent Study - (3 days)**

This week you’ll have the opportunity to do another 3-day independent study on a topic of your choice using your newly acquired Python skills. The goal is to give you time to select data, explore it looking for an interesting angle, create visualizations, and then present your findings to the class and FSA staff in a 10 minute solo presentation.

The product of the week’s efforts should be something that you can share with potential employers via LinkedIn, Github, your own blog/webpage, etc.

A very rough outline for the week is as follows:

* Tuesday (3/21)- Dataset selection, data cleaning and exploration via visualization, descriptive statistics, etc.
* Wednesday (3/22) - Analyze, discover, and outline your ‘story’
* Thursday (3/23) - Draft presentation, decide what is in, what is out. Which visualizations help you tell your story in 10 minutes. Prepare your final presentation, share it with your colleagues if you’d like, practice your presentation.
* Friday (3/24) PM - present to staff and students.

| Data | Source | Description | Notes |
| --- | --- | --- | --- |
| AirBnB data | <http://insideairbnb.com/get-the-data> | Quarterly data for the last year for each region is available | Pick a city! |
| NYC Open Data | <https://opendata.cityofnewyork.us/> | Open Data is free public data published by New York City agencies and other partners | Choose carefully. Some of these datasets are very large and require extensive cleaning |
| Baseball data | [seanlahman.com](https://www.seanlahman.com/baseball-archive/statistics) | Individual/team, batting, pitching, data available since the dawn of time | Take me out to the ballgame… with data. |

Standup groups – each day during the project period we will meet in standup groups to discuss any progress blockers, challenges, and hurdles that you might be experiencing. These are the groups with your mentor/advisor:

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

## You are welcome to meet with Monica or Vadim or Dennis, and we will all have expanded office hours this week. Some precautionary advice:

* Choose your dataset by noon on Tuesday. Don’t fall prey to dataset vagabonding.
* Don’t try to boil the ocean (you only have 3 days).
* Don’t start with a conclusion (i.e. confirmation bias). Instead, allow the data to lead you to a conclusion.
* Avoid fixating on one task or part of the analysis. Data cleaning for 2.5 days only leaves ½ day for everything else.
* Visualize early to understand what kind of data you are dealing with. Don’t assume that visualizations are just the end product.
* Feel free to drop columns that don’t contribute to your analysis. You can always go back and get them if needed later on.
* Present in Jupyter or Powerpoint. Both have advantages and disadvantages.
* If you want to try machine learning, make sure you identify multiple input features and one outcome/classification variable. Not all datasets are well suited to machine learning.