**Dataset Description**

# **HOUSING NEW YORK UNITS**

# 

## **TASK**

One dataset will be provided to each team. The dataset can be found in the folder corresponding to the team number.

The team’s task is to understand the dataset completely, and perform exploratory data analysis on it. This would entail using knowledge acquired from the class sessions to plot visualizations which would help understand the dataset better, and derive insights from it. These visualizations and basic insights are a must for the project. However, the focus of this project is to augment theoretical learning of the class with invaluable hands-on knowledge.

This project is also meant to increase awareness and encourage exploration of fields and topics which are overlapping with the field of Data Science. Thus, a portion of the project is meant to test the team’s willingness to learn and explore about new topics and techniques. Take note, however, that these might be topics which are not part of the Data Science course, but nonetheless will impart a great deal of knowledge.

To aid the thinking process, there is a section titled ‘Possible tracks for insights’ which is meant to provide an example of a possible direction to begin thinking in. The questions provided in that section are NOT a checklist of things to be done, as the basic visualizations need to be implemented, even though they are not mentioned in the section. It could also be the case that the question provided in this section does not provide any insight. In such cases, simply consider it as a direction to be thinking in, and feel free to skip the results of the question in the report. Marks will be awarded for any and all techniques not part of the course which have been implemented.

# 

# **DESCRIPTION**

A dataset is a collection of data. In a dataset, each row is a particular instance or record of an event, and each column corresponds to one variable or feature.

The Department of Housing Preservation and Development (HPD) reports on buildings, units, and projects that began after January 1, 2014 and are counted towards the Housing New York plan. The Housing New York Units by Building file presents this data by building, and includes building-level data, such as house number, street name, BBL, and BIN for each building in a project. The unit counts are provided by building. This dataset presents the information about the city of New York

**Take note**, however, that the data may have to be cleaned, with a number of erroneous or blank values, with a number of outliers. The dataset will be a CSV file or a number of them. Your task is to clean the dataset, and perform analysis on it, plot relevant visualizations and to gain some meaningful insight into the data using the Data Science skills you have acquired. If you deem any columns to be irrelevant to your analysis, you can discard them. However, you will have to report clearly why this column was discarded.

## **COLUMN EXPLANATION**

In this section, a short description of the meaning of each column can be found. Along with this, the type of data in the column is indicated.

1. Project ID: The Project ID is a unique numeric identifier assigned to each project by HPD.
2. Project Name: The Project Name is the name assigned to the project by HPD.
3. Project Start Date: The Project Start Date is the date of the project loan or agreement closing.
4. Project Completion Date: The Project Completion Date is the date that the last building in the project was completed. If the project has not yet completed, then the field is blank.
5. Building ID: The Building ID is a unique numeric identifier assigned to each building by HPD.
6. Number: The House Number is the street number in the building’s address. E.g., the house number is ‘100’ in ‘100 Gold Street.’
7. Street: The Street Name is the name of the street in the building’s address. E.g., the street name is ‘Gold Street’ in ‘100 Gold Street.’
8. Borough: The Borough is the borough where the building is located.
9. Postcode: Zip code
10. BBL: The BBL (Borough, Block, and Lot) is a unique identifier for each tax lot in the City.
11. BIN: The BIN (Building Identification Number) is a unique identifier for each building in the City.
12. Community Board: The Community Board field indicates the New York City Community District where the building is located.
13. Council District: The Council District indicates the New York City Council District where the building is located.
14. Census Tract: The Census Tract indicates the 2010 U.S. Census Tract where the building is located.
15. NTA - Neighborhood Tabulation Area: The Neighborhood Tabulation Area indicates the New York City Neighborhood Tabulation Area where the building is located.
16. Latitude: The Latitude and Longitude specify the location of the property on the earth’s surface. The coordinates provided are an estimate of the location based on the street segment and address range.
17. Longitude: The Latitude and Longitude specify the location of the property on the earth’s surface. The coordinates provided are an estimate of the location based on the street segment and address range.
18. Latitude (Internal): The Latitude (Internal) and Longitude (Internal) specify the location of the property on the earth’s surface. The coordinates provided are of the internal centroid derived from the tax lot.
19. Longitude (Internal): The Latitude (Internal) and Longitude (Internal) specify the location of the property on the earth’s surface. The coordinates provided are of the internal centroid derived from the tax lot.
20. Building Completion Date: The Building Completion Date is the date the building was completed. The field is blank if the building has not completed.
21. Reporting Construction Type: The Reporting Construction Type field indicates whether the building is categorized as ‘new construction’ or ‘preservation’ in Housing New York statistics. Note that some preservation projects included here may not actually involve construction, because they extend the project’s regulatory restrictions but do not require rehabilitation.
22. Extended Affordability Only: The Extended Affordability Only field indicates whether the project is considered to be Extended Affordability. An extended affordability project involves no construction, but secures an extended or new regulatory agreement. All extended affordability projects have a ‘reporting construction type’ of ‘preservation.’
23. Prevailing Wage Status: The Prevailing Wage Status field indicates whether the project is subject to prevailing wage requirements, such as Davis Bacon.
24. Extremely Low Income Units: Extremely Low Income Units are units with rents that are affordable to households earning 0 to 30% of the area median income (AMI).
25. Very Low Income Units: Very Low Income Units are units with rents that are affordable to households earning 31 to 50% of the area median income (AMI).
26. Low Income Units: Low Income Units are units with rents that are affordable to households earning 51 to 80% of the area median income (AMI).
27. Moderate Income Units: Moderate Income Units are units with rents that are affordable to households earning 81 to 120% of the area median income (AMI).
28. Middle Income Units: Middle Income Units are units with rents that are affordable to households earning 121 to 165% of the area median income (AMI).
29. Other Income Units: Other Units are units reserved for building superintendents.
30. Studio Units: Studio Units are units with 0-bedrooms.
31. 1-BR Units: 1-BR Units are units with 1-bedroom.
32. 2-BR Units: 2-BR Units are units with 2-bedrooms.
33. 3-BR Units: 3-BR Units are units with 3-bedrooms.
34. 4-BR Units: 4-BR Units are units with 4-bedrooms.
35. 5-BR Units: 5-BR Units are units with 5-bedrooms.
36. 6-BR+ Units: 6-BR+ Units are units with 6-bedrooms or more.
37. Unknown-BR Units: Unknown-BR Units are units with an unknown number of bedrooms.
38. Counted Rental Units: Counted Rental Units are the units in the building, counted toward the Housing New York plan, where assistance has been provided to landlords in exchange for a requirement for affordable units.
39. Counted Homeownership Units: Counted Homeownership Units are the units in the building, counted toward the Housing New York Plan, where assistance has been provided directly to homeowners.
40. All Counted Units: The Counted Units field indicates the total number of affordable units, counted towards the Housing New York plan, that are in the building.
41. Total Units: The Total Units field indicates the total number of units, affordable and market rate, in each building.

## **POSSIBLE TRACKS FOR INSIGHTS**

In this section, a small number of possible insights by manipulating and analysing the data are presented. This is certainly **NOT** a checklist of things to do in your project, but rather aims to provide a generic example of the direction to begin thinking in. It could certainly be the case that more valuable insights may be possible, which are not listed here. Some suggestions may be beyond the syllabus and are included only for learning, so please feel free to go out and learn how to do them.

Which is the borough with the most 1-BR units?

Is there a correlation between the number of rooms and the category of income to which the units belong, for a particular area?

Which is the borough with the most low income units?

## **POTENTIALLY USEFUL LINKS**

In this section, you will find a number of links, which can be used to brush up your knowledge or help resolve issues.

### Python Tutorial

### <https://www.learnpython.org/>

### Conda Cheatsheet

<https://docs.conda.io/projects/conda/en/4.6.0/_downloads/52a95608c49671267e40c689e0bc00ca/conda-cheatsheet.pdf>

### Pandas Documentation

<https://pandas.pydata.org/pandas-docs/stable/>

### Pandas Quickstart (Short Tutorial)

<https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html>

### NumPy Quickstart (Short Tutorial)

<https://docs.scipy.org/doc/numpy/user/quickstart.html>

### Seaborn Documentation

<https://seaborn.pydata.org/introduction.html>

### Seaborn In-Depth Tutorial

<https://seaborn.pydata.org/tutorial.html#tutorial>

### Matplotlib Tutorial

<https://matplotlib.org/3.1.1/tutorials/index.html>

### Scikit-Learn Tutorial

<https://scikit-learn.org/stable/tutorial/index.html>