

# Assignment2\_DATA608

September 4, 2022

```
[ ]: # ! which python
      # ! pip install datashader
      # pip install --upgrade xarray==2022.3.0
      # ! pip install pyproj
      # ! pip install colorlover
      # ! pip install plotly
      # jupyter nbextension enable --py widgetsnbextension --sys-prefix
```

```
[111]: import datashader as ds
import datashader.transfer_functions as tf
import datashader.glyphs
from datashader import reductions
from datashader.core import bypixel
from datashader.utils import lnglat_to_meters as webm, export_image
from datashader.colors import colormap_select, Greys9, viridis, inferno
import copy

from pyproj import Proj, transform
import numpy as np
import pandas as pd
import urllib
import json
import datetime
import colorlover as cl

import plotly.offline as py
import plotly.graph_objs as go
from plotly import tools

# from shapely.geometry import Point, Polygon, shape
# In order to get shapely, you'll need to run [pip install shapely.geometry]
↳ from your terminal

from functools import partial

from IPython.display import GeoJSON
```

```
py.init_notebook_mode()
```

For module 2 we'll be looking at techniques for dealing with big data. In particular binning strategies and the datashader library (which possibly proves we'll never need to bin large data for visualization ever again.)

To demonstrate these concepts we'll be looking at the PLUTO dataset put out by New York City's department of city planning. PLUTO contains data about every tax lot in New York City.

PLUTO data can be downloaded from [here](#). Unzip them to the same directory as this notebook, and you should be able to read them in using this (or very similar) code. Also take note of the data dictionary, it'll come in handy for this assignment.

```
[112]: # Code to read in v17, column names have been updated (without upper case
↳ letters) for v18

# bk = pd.read_csv('PLUTO17v1.1/BK2017V11.csv')
# bx = pd.read_csv('PLUTO17v1.1/BX2017V11.csv')
# mn = pd.read_csv('PLUTO17v1.1/MN2017V11.csv')
# qn = pd.read_csv('PLUTO17v1.1/QN2017V11.csv')
# si = pd.read_csv('PLUTO17v1.1/SI2017V11.csv')

# ny = pd.concat([bk, bx, mn, qn, si], ignore_index=True)

ny = pd.read_csv('/Users/tponnada/Downloads/pluto_22v2.csv')

# Getting rid of some outliers
ny = ny[(ny['yearbuilt'] > 1850) & (ny['yearbuilt'] < 2020) & (ny['numfloors'] !
↳ = 0)]
```

I'll also do some prep for the geographic component of this data, which we'll be relying on for datashader.

You're not required to know how I'm retrieving the latitude and longitude here, but for those interested: this dataset uses a flat x-y projection (assuming for a small enough area that the world is flat for easier calculations), and this needs to be projected back to traditional latitude and longitude.

```
[123]: # wgs84 = Proj("+proj=longlat +ellps=GRS80 +datum=NAD83 +no_defs")
# nyli = Proj("+proj=lcc +lat_1=40.66666666666666 +lat_2=41.03333333333333
↳ +lat_0=40.16666666666666 +lon_0=-74 +x_0=300000 +y_0=0 +ellps=GRS80
↳ +datum=NAD83 +to_meter=0.3048006096012192 +no_defs")
# ny['xcoord'] = 0.3048*ny['xcoord']
# ny['ycoord'] = 0.3048*ny['ycoord']
# ny['lon'], ny['lat'] = transform(nyli, wgs84, ny['xcoord'].values,
↳ ny['ycoord'].values)
```

```
# ny = ny[(ny['lon'] < -60) & (ny['lon'] > -100) & (ny['lat'] < 60) &
↳ (ny['lat'] > 20)]

#Defining some helper functions for DataShader
background = "black"
export = partial(export_image, background = background, export_path = "export")
cm = partial(colormap_select, reverse =(background != "black"))
```

Part 1: Binning and Aggregation Binning is a common strategy for visualizing large datasets. Binning is inherent to a few types of visualizations, such as histograms and 2D histograms (also check out their close relatives: 2D density plots and the more general form: heatmaps).

While these visualization types explicitly include binning, any type of visualization used with aggregated data can be looked at in the same way. For example, lets say we wanted to look at building construction over time. This would be best viewed as a line graph, but we can still think of our results as being binned by year:

```
[124]: trace = go.Scatter(
    # I'm choosing BBL here because I know it's a unique key.
    x = ny.groupby('yearbuilt').count()['bbl'].index,
    y = ny.groupby('yearbuilt').count()['bbl']
)

layout = go.Layout(
    xaxis = dict(title = 'Year Built'),
    yaxis = dict(title = 'Number of Lots Built')
)

fig = go.FigureWidget(data = [trace], layout = layout)

fig.show()
```

Something looks off... You're going to have to deal with this imperfect data to answer this first question.

But first: some notes on pandas. Pandas dataframes are a different beast than R dataframes, here are some tips to help you get up to speed:

Hello all, here are some pandas tips to help you guys through this homework:

Indexing and Selecting: .loc and .iloc are the analogs for base R subsetting, or filter() in dplyr

Group By: This is the pandas analog to group\_by() and the appended function the analog to summarize(). Try out a few examples of this, and display the results in Jupyter. Take note of what's happening to the indexes, you'll notice that they'll become hierarchical. I personally find this more of a burden than a help, and this sort of hierarchical indexing leads to a fundamentally different experience compared to R dataframes. Once you perform an aggregation, try running the resulting hierarchical dataframe through a reset\_index().

Reset\_index: I personally find the hierarchical indexes more of a burden than a help, and this sort of hierarchical indexing leads to a fundamentally different experience compared to R dataframes.

`reset_index()` is a way of restoring a dataframe to a flatter index style. Grouping is where you'll notice it the most, but it's also useful when you filter data, and in a few other split-apply-combine workflows. With pandas indexes are more meaningful, so use this if you start getting unexpected results.

Indexes are more important in Pandas than in R. If you delve deeper into the using python for data science, you'll begin to see the benefits in many places (despite the personal gripes I highlighted above.) One place these indexes come in handy is with time series data. The pandas docs have a huge section on datetime indexing. In particular, check out `resample`, which provides time series specific aggregation.

Merging, joining, and concatenation: There's some overlap between these different types of merges, so use this as your guide. `Concat` is a single function that replaces `cbind` and `rbind` in R, and the results are driven by the indexes. Read through these examples to get a feel on how these are performed, but you will have to manage your indexes when you're using these functions. Merges are fairly similar to merges in R, similarly mapping to SQL joins.

Apply: This is explained in the "group by" section linked above. These are your analogs to the `plyr` library in R. Take note of the lambda syntax used here, these are anonymous functions in python. Rather than predefining a custom function, you can just define it inline using `lambda`.

Browse through the other sections for some other specifics, in particular reshaping and categorical data (pandas' answer to factors.) Pandas can take a while to get used to, but it is a pretty strong framework that makes more advanced functions easier once you get used to it. Rolling functions for example follow logically from the apply workflow (and led to the best google results ever when I first tried to find this out and googled "pandas rolling")

Google Wes Mckinney's book "Python for Data Analysis," which is a cookbook style intro to pandas. It's an O'Reilly book that should be pretty available out there.

Question:

After a few building collapses, the City of New York is going to begin investigating older buildings for safety. The city is particularly worried about buildings that were unusually tall when they were built, since best-practices for safety hadn't yet been determined. Create a graph that shows how many buildings of a certain number of floors were built in each year (note: you may want to use a log scale for the number of buildings).

Find a strategy to bin buildings (It should be clear 20-29-story buildings, 30-39-story buildings, and 40-49-story buildings were first built in large numbers, but does it make sense to continue in this way as you get taller?)

```
[125]: # Start your answer here, inserting more cells as you go along

#Run a few summary statistics to get a better understanding of the data set

ny['yearbuilt'].describe()
```

```
[125]: count    812064.000000
      mean      1941.167871
      std       30.580862
```

```

min          1851.000000
25%          1920.000000
50%          1931.000000
75%          1960.000000
max          2019.000000
Name: yearbuilt, dtype: float64

```

```
[126]: ny['numfloors'].describe()
```

```

[126]: count      809793.000000
      mean         2.459627
      std         1.947155
      min         1.000000
      25%         2.000000
      50%         2.000000
      75%         2.750000
      max         104.000000
      Name: numfloors, dtype: float64

```

Since the mean is greater than the median (50th percentile), this indicates that the distribution likely has a positive skew, i.e. there are more buildings of modern vintage and that there are a few buildings which are really high (skyscrapers) that are pulling the mean > median.

```
[127]: #Check the distribution of buildings and floor numbers by year built.
```

```

year_built1 = ny['yearbuilt'].sample(n = 10)

for i in year_built1:

    trace0 = go.Histogram(

        x = ny['numfloors'][ny['yearbuilt'] == i], nbinsx = 10)

fig = tools.make_subplots(rows = 1, cols = 1)

fig.append_trace(trace0, 1, 1)

py.ipplot(fig, filename = 'Bin 10 graph')

```

```
[128]: #Check the distribution of buildings and floor numbers by year built.
```

```

year_built2 = ny['yearbuilt'].sample(n = 10)

for i in year_built2:

    trace0 = go.Histogram(

```

```

x = ny['numfloors'][ny['yearbuilt'] == i], nbinsx = 20)

fig = tools.make_subplots(rows = 1, cols = 1)

fig.append_trace(trace0, 1, 1)

py.ipplot(fig, filename = 'Bin 20 graph')

```

The above graphic suggests a bin size of 20 is perhaps more appropriate.

[129]: *# The following code was borrowed from Lidia25 and altered to show a better  
→ representation of the bins.*

```

import warnings
warnings.filterwarnings('ignore')
yfdf = ny[['yearbuilt', 'numfloors']]
yfdf['y10'] = (yfdf['yearbuilt'] // 10 * 10).astype(int)
bins = (20, 30, 40, 50, 70, 210)
cdf = yfdf.groupby(['y10', pd.cut(yfdf['numfloors'], bins)]) \
    .count() \
    .drop(['numfloors'], axis = 1) \
    .fillna(0) \
    .reset_index()

```

[130]:

```

import plotly.graph_objects as go
groups = cdf.groupby('numfloors')
fig = go.Figure()
for g in groups.groups:
    group = groups.get_group(g)
    fig.add_trace(go.Bar(x = group['y10'], y = group['yearbuilt'], name =
→ str(g)))
fig.update_layout(barmode = 'stack', xaxis = {'categoryorder': 'category
→ ascending'})
fig.show()

```

Since we studied the dataset above and determined that there are a few buildings with really high floors, we decide to eliminate the lower limit of 0 for bin size. Furthermore, using the bin size of 20 appears to be optimal till we reach a threshold of 70 floors after which the bins are collapsed to show depth (as there are fewer buildings beyond 70 floors).

A quick look at the chart shows that buildings are getting taller over time. While there were ~110 buildings with 30 floors in the 1920's, that number subsequently increased to 175 in the 1960's. Since the 60's however, the number of floors have increased with 30-40 floors being a popular choice (with the exception of the 90's period which overall saw relatively little construction, could be because of the S&L and/or CRE crisis). Furthermore, since the 80's (again with the exception of the 90's which saw relatively little construction overall), the 40-50 as well as the 50-70 floor high buildings seem to be gaining popularity. One reason could be that building technology has improved over time, while it may be true that older vintage taller buildings might be more susceptible to

safety issues etc. due to a lack of adherence to modern building codes, there seems to be greater confidence with building taller buildings over time due to technology.

Part 2: Datashader Datashader is a library from Anaconda that does away with the need for binning data. It takes in all of your datapoints, and based on the canvas and range returns a pixel-by-pixel calculations to come up with the best representation of the data. In short, this completely eliminates the need for binning your data.

As an example, lets continue with our question above and look at a 2D histogram of YearBuilt vs NumFloors:

```
[131]: fig = go.FigureWidget(
        data = [
            go.Histogram2d(x = ny['yearbuilt'], y = ny['numfloors'], autobin =
→False, ybins = {'size': 1}, colorscale = 'Greens')
        ]
    )

    fig.show()
```

This shows us the distribution, but it's subject to some biases discussed in the Anaconda notebook Plotting Perils.

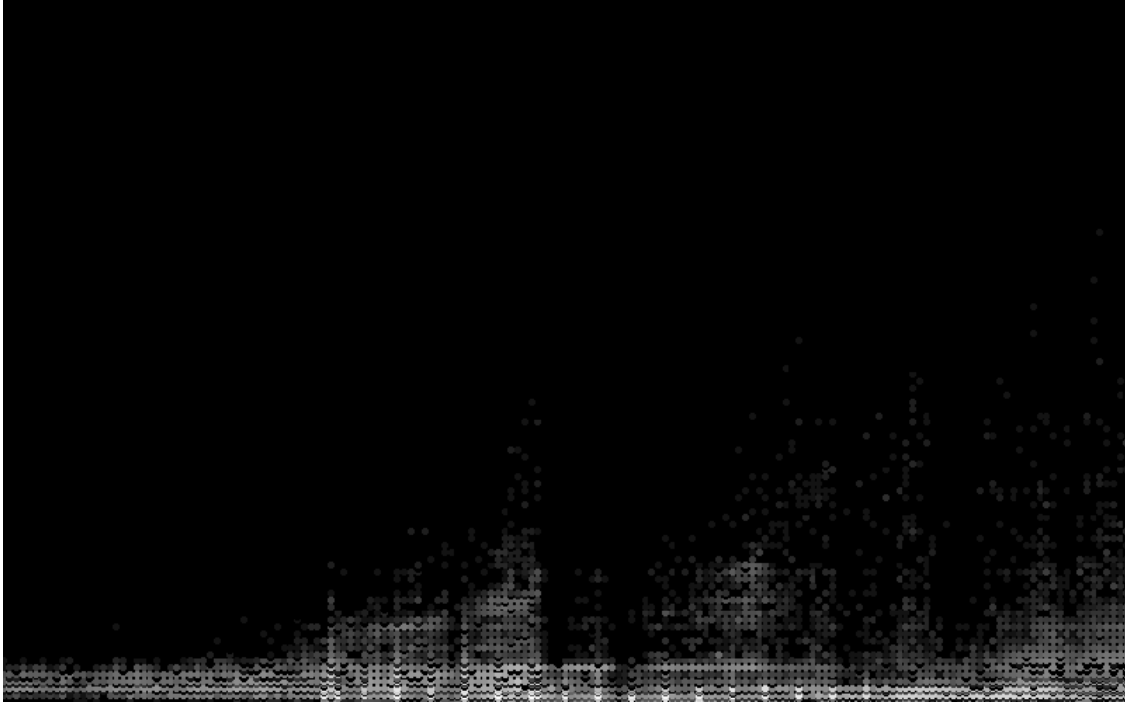
Here is what the same plot would look like in datashader:

```
[132]: #Defining some helper functions for DataShader

background = "black"
export = partial(export_image, background = background, export_path = "export")
cm = partial(colormap_select, reverse=(background != "black"))

cvs = ds.Canvas(800, 500, x_range = (ny['yearbuilt'].min(), ny['yearbuilt'].
→max()),
                y_range = (ny['numfloors'].min(),
→ny['numfloors'].max()))
agg = cvs.points(ny, 'yearbuilt', 'numfloors')
view = tf.shade(agg, cmap = cm(Greys9), how = 'log')
export(tf.spread(view, px = 2), 'yearvsnumfloors')
```

[132]:



That's technically just a scatterplot, but the points are smartly placed and colored to mimic what one gets in a heatmap. Based on the pixel size, it will either display individual points, or will color the points of denser regions.

Datashader really shines when looking at geographic information. Here are the latitudes and longitudes of our dataset plotted out, giving us a map of the city colored by density of structures:

```
[133]: NewYorkCity = (( 913164.0, 1067279.0), (120966.0, 272275.0))
cvs = ds.Canvas(700, 700, *NewYorkCity)
agg = cvs.points(ny, 'xcoord', 'ycoord')
view = tf.shade(agg, cmap = cm.inferno, how = 'log')
export(tf.spread(view, px = 2), 'firery')
```

[133]:





Interestingly, since we're looking at structures, the large buildings of Manhattan show up as less dense on the map. The densest areas measured by number of lots would be single or multi family townhomes.

Unfortunately, Datashader doesn't have the best documentation. Browse through the examples from their github repo. I would focus on the visualization pipeline and the US Census Example for the question below. Feel free to use my samples as templates as well when you work on this problem.

Question:

You work for a real estate developer and are researching underbuilt areas of the city. After looking in the Pluto data dictionary, you've discovered that all tax assessments consist of two parts: The assessment of the land and assessment of the structure. You reason that there should be a correlation between these two values: more valuable land will have more valuable structures on them

(more valuable in this case refers not just to a mansion vs a bungalow, but an apartment tower vs a single family home). Deviations from the norm could represent underbuilt or overbuilt areas of the city. You also recently read a really cool blog post about bivariate choropleth maps, and think the technique could be used for this problem.

Datashader is really cool, but it's not that great at labeling your visualization. Don't worry about providing a legend, but provide a quick explanation as to which areas of the city are overbuilt, which areas are underbuilt, and which areas are built in a way that's properly correlated with their land value.

```
[134]: # This part of the code was borrowed from Lidia25 and altered to align with a
        ↪ better solution.

        #Assessment of the structure column was missing in Pluto dataset which has been
        ↪ calculated by subtracting assessland from assesstot to arrive at the
        ↪ structure value.

ny['assesstructure'] = ny['assesstot'] - ny['assessland']

asdf = pd.DataFrame(ny[['assessland', 'assesstructure', 'latitude',
        ↪ 'longitude']])
asdf['land_segment'] = 'med'
asdf.loc[asdf['assessland'] < 8000, 'land_segment'] = 'low'
asdf.loc[asdf['assessland'] >= 12000, 'land_segment'] = 'high'
asdf['struct_segment'] = 'med'
asdf.loc[asdf['assesstructure'] < 17500, 'struct_segment'] = 'low'
asdf.loc[asdf['assesstructure'] >= 27500, 'struct_segment'] = 'high'
asdf['segment'] = pd.Categorical(asdf['land_segment'] + '-' +
        ↪ asdf['struct_segment'])
asdf.describe()
```

```
[134]:
```

	assessland	assesstructure	latitude	longitude
count	8.120640e+05	8.120640e+05	811915.000000	811915.000000
mean	1.128162e+05	4.144469e+05	40.691828	-73.920210
std	4.174929e+06	7.558537e+06	0.083734	0.117301
min	0.000000e+00	0.000000e+00	40.498445	-74.255628
25%	1.038000e+04	2.952000e+04	40.627946	-73.980437
50%	1.416000e+04	4.224000e+04	40.685589	-73.910572
75%	2.196000e+04	7.836000e+04	40.745247	-73.836626
max	3.205634e+09	4.343287e+09	40.913967	-73.700381

```
[135]: # Plot the data

NewYorkCity = ((-74.255628, -73.700381), (40.498445, 40.913967))
cvs = ds.Canvas(700, 700, *NewYorkCity)
agg = cvs.points(asdf, 'longitude', 'latitude', ds.count_cat('segment'))
view = tf.shade(agg, cmap = cm.inferno)
export(tf.spread(view, px = 2), 'firery')
```

[135] :



From the graphic above, Manhattan and central areas appear to be the most overbuilt with the highest assessed land values. Some areas in Brooklyn and Queens have higher density as well with outlying areas (Airport, Brooklyn/Queens border) appearing to have a low-to-medium density. Assessed total and land values decline as we move away from Manhattan which aligns with intuition.

[ ] :