In [1]:

```
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.plotly as py
import plotly.graph_objs as go

from plotly import tools
from plotly import __version__

import cufflinks as cf
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot

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

from functools import partial

from IPython.display import GeoJSON

import plotly
plotly.offline.init_notebook_mode()
```

```
c:\users\sergi\pycharmprojects\dnns\venv\lib\site-packages\datashader\transfer_functions.
py:21: FutureWarning: xarray subclass Image should explicitly define __slots__
  class Image(xr.DataArray):
```

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.

In [2]:

```
# Code to read in v17, column names have been updated (without upper case letters) for v1
8

#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('pluto_19v1.csv')


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

```
c:\users\sergi\pycharmprojects\dnns\venv\lib\site-packages\IPython\core\interactiveshell.
py:3058: DtypeWarning:

Columns (16,17,18,20,22) have mixed types. Specify dtype option on import or set low_memo
ry=False.
```

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 lattitude 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 lattitude and longitude.

In [3]:

```
wgs84 = Proj("+proj=longlat +ellps=GRS80 +datum=NAD83 +no_defs")
nyli = Proj("+proj=lcc +lat_1=40.66666666666666 +lat_2=41.03333333333333 +lat_0=40.166666
66666666 +lon_0=-74 +x_0=300000 +y_0=0 +ellps=GRS80 +datum=NAD83 +to_meter=0.304800609601
2192 +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:

In [4]:

```
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.Figure(data = [trace], layout = layout)

py.iplot(fig)
```

```
c:\users\sergi\pycharmprojects\dnns\venv\lib\site-packages\IPython\core\display.py:694: U
serWarning:

Consider using IPython.display.IFrame instead
```

Out[4]:

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 datafrome 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 unusally 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?)

In [5]:

```
# Start your answer here, inserting more cells as you go along


# I started with some initial EDA (exploratory data analysis) to determine what bins to u
se for the data
# I created a dataframe with yearbuilt, then Count of numfloors, Min of numfloors, Max of
numfloors,
# Mean of numfloors, Standard Deviation of numfloors, and then a CutOff value which is Me
an + 4 times the Std Dev

c1 = ny.groupby('yearbuilt').count()['numfloors']
c2 = ny.groupby('yearbuilt').min()['numfloors']
c3 = ny.groupby('yearbuilt').max()['numfloors']
c4 = ny.groupby('yearbuilt').mean()['numfloors']
c5 = ny.groupby('yearbuilt').std()['numfloors']
c6 = c4 + (4 * c5)

nynf = {
'Cnt_Flrs': c1,
'Min_Flrs': c2,
'Max_Flrs': c3,
'Mean_Flrs': c4,
'Stddev_Flrs': c5,
'CutOff': c6}

nynf_df = pd.DataFrame(nynf)
```

In [6]:

```
# Looking at this dataframe, we see the lowest numfloors value is 1 and the max is 119
# Particularly in the last 15 years, you regularly see values above 79
# As such, I chose to set the bin ranges as 0-19, 20-29, 30-39, 40-49, 50-69, 70-94, and
95-119
# To make the data frame more manageable, I created a new dataframe which is 3 columns fr
om the original

ny_selected = ny[['yearbuilt','bbl','numfloors']].copy()

# I then designed a function which is in turn used to populate a bin column

def func(row):
    if row['numfloors'] >= 20.0 and row['numfloors'] <= 29.0:
```

```
        return '20-29 floors'
    elif row['numfloors'] >= 30.0 and row['numfloors'] <= 39.0:
        return '30-39 floors'
    elif row['numfloors'] >= 40.0 and row['numfloors'] <= 49.0:
        return '40-49 floors'
    elif row['numfloors'] >= 50.0 and row['numfloors'] <= 69.0:
        return '50-69 floors'
    elif row['numfloors'] >= 70.0 and row['numfloors'] <= 94.0:
        return '70-94 floors'
    elif row['numfloors'] >= 95.0 and row['numfloors'] <= 119.0:
        return '95-119 floors'
    else:
        return '0-19 floors'

ny_selected['bin'] = ny_selected.apply(func, axis=1)

# Then I created a plot showing counts by bin

range1 = ny_selected.loc[ny_selected["bin"] == '0-19 floors']
range2 = ny_selected.loc[ny_selected["bin"] == '20-29 floors']
range3 = ny_selected.loc[ny_selected["bin"] == '30-39 floors']
range4 = ny_selected.loc[ny_selected["bin"] == '40-49 floors']
range5 = ny_selected.loc[ny_selected["bin"] == '50-69 floors']
range6 = ny_selected.loc[ny_selected["bin"] == '70-94 floors']
range7 = ny_selected.loc[ny_selected["bin"] == '95-119 floors']

bin2 = go.Scatter(
    x = range2.groupby('yearbuilt').count()['numfloors'].index,
    y = range2.groupby('yearbuilt').count()['numfloors'],
    name = "20-29 Floors"
)

bin3 = go.Scatter(
    x = range3.groupby('yearbuilt').count()['numfloors'].index,
    y = range3.groupby('yearbuilt').count()['numfloors'],
    name = "30-39 Floors"
)

bin4 = go.Scatter(
    x = range4.groupby('yearbuilt').count()['numfloors'].index,
    y = range4.groupby('yearbuilt').count()['numfloors'],
    name = "40-49 Floors"
)

bin5 = go.Scatter(
    x = range5.groupby('yearbuilt').count()['numfloors'].index,
    y = range5.groupby('yearbuilt').count()['numfloors'],
    name = "50-69 Floors"
)

bin6 = go.Scatter(
    x = range6.groupby('yearbuilt').count()['numfloors'].index,
    y = range6.groupby('yearbuilt').count()['numfloors'],
    name = "70-94 Floors"
)

bin7 = go.Scatter(
    x = range7.groupby('yearbuilt').count()['numfloors'].index,
    y = range7.groupby('yearbuilt').count()['numfloors'],
    name = "95-119 Floors"
)


layout2 = go.Layout(
    xaxis = dict(title = 'Year Built'),
    yaxis = dict(title = 'Number of Buildings')
)
```

In [7]:

```
#Graph options and Data
```

```
bins = [bin2, bin3, bin4, bin5, bin6, bin7]

fig2 = go.Figure(data = bins, layout = layout2)
py.iplot(fig2, filename = 'ny-year-built2')
```

Out[7]:

In [8]:

```
# One area of interest for NYC is "buildings that were unusually tall when they were buil
t". I merged the nynf_df I
# created to the original ny dataframe and then subsetted to find properties which were m
ore than 4 standard deviations
# above the mean for the year (omitted years where less than 20 buildings were built). DF
is called "ny_outliers".

ny_outliers_base = pd.merge(ny_selected, nynf_df, on='yearbuilt', how='left')
ny_outliers = ny_outliers_base[(ny_outliers_base['Cnt_Flrs'] >= 20) & (ny_outliers_base[
'numfloors'] > ny_outliers_base['CutOff'])]
```

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:

In [9]:

```
yearbins = 200
floorbins = 200

yearBuiltCut = pd.cut(ny['yearbuilt'], np.linspace(ny['yearbuilt'].min(), ny['yearbuilt'
].max(), yearbins))
numFloorsCut = pd.cut(ny['numfloors'], np.logspace(1, np.log(ny['numfloors'].max()), flo
```

```
orbins))

xlabels = np.floor(np.linspace(ny['yearbuilt'].min(), ny['yearbuilt'].max(), yearbins))
ylabels = np.floor(np.logspace(1, np.log(ny['numfloors'].max()), floorbins))

data = [
    go.Heatmap(z = ny.groupby([numFloorsCut, yearBuiltCut])['bbl'].count().unstack().fil
lna(0).values,
                colorscale = 'Greens', x = xlabels, y = ylabels)
]

py.iplot(data)
```

Out[9]:

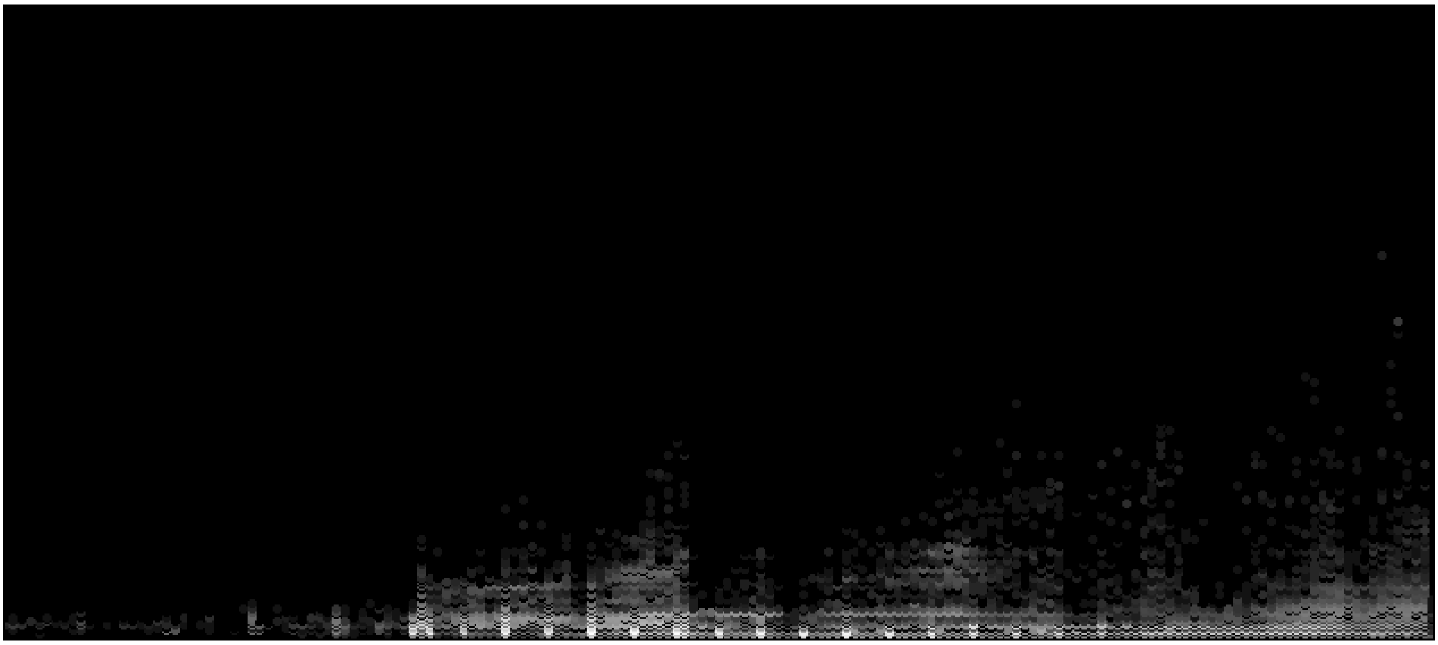
This shows us the distribution, but it's subject to some biases discussed in the Anaconda notebook [Plotting Perils](#).

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

In [10]:

```
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')
```

Out[10]:



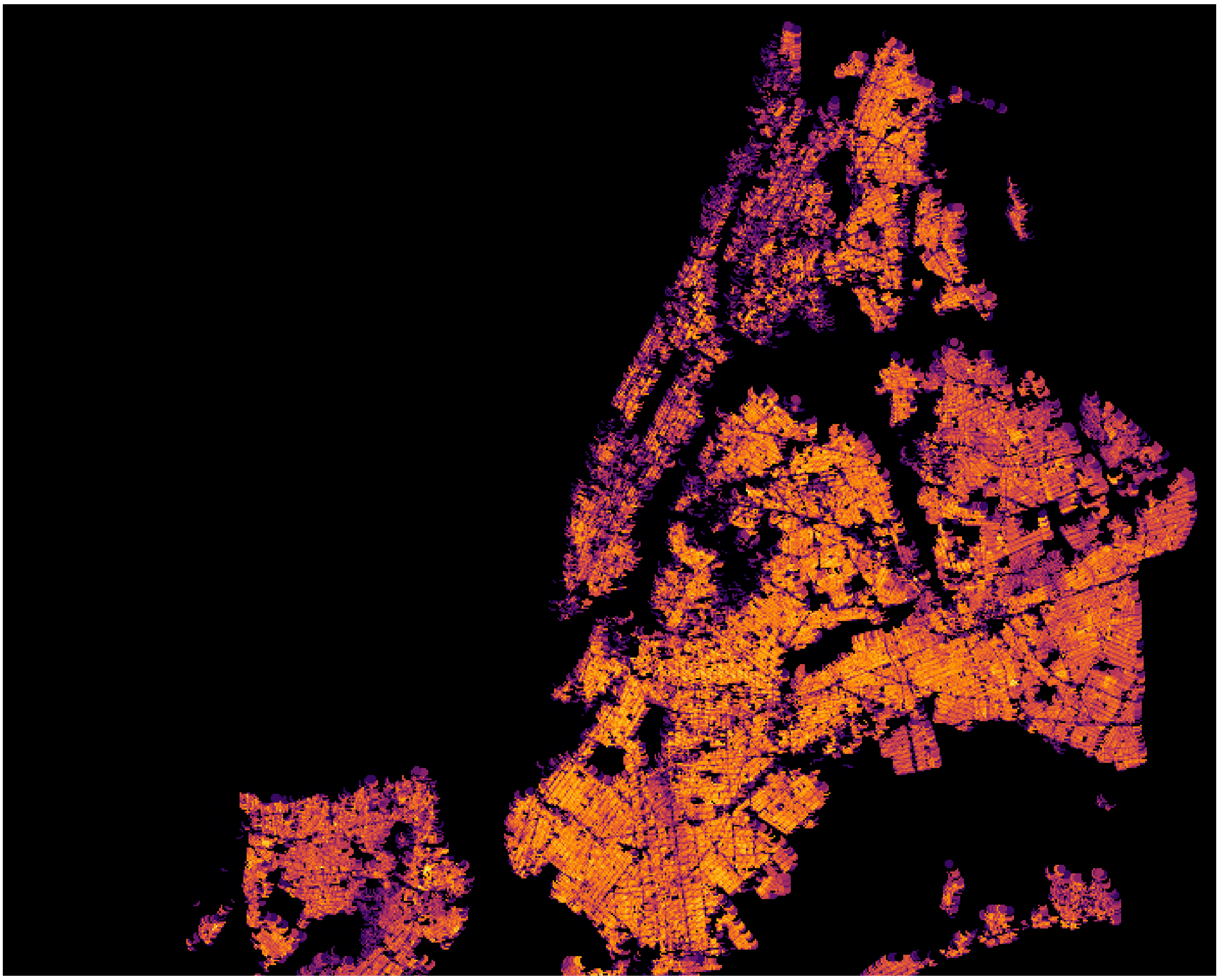
That's technically just a scatterplot, but the points are smartly placed and colored to mimic what one gets in a heatmap. Based on the pixel size, it will either display individual points, or will color the points of denser regions.

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:

In [11]:

```
NewYorkCity   = (( -74.29,  -73.69), (40.49, 40.92))
cvs = ds.Canvas(700, 700, *NewYorkCity)
agg = cvs.points(ny, 'lon', 'lat')
view = tf.shade(agg, cmap = cm(inferno), how='log')
export(tf.spread(view, px=2), 'firery')
```

Out[11]:



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.

In [12]:

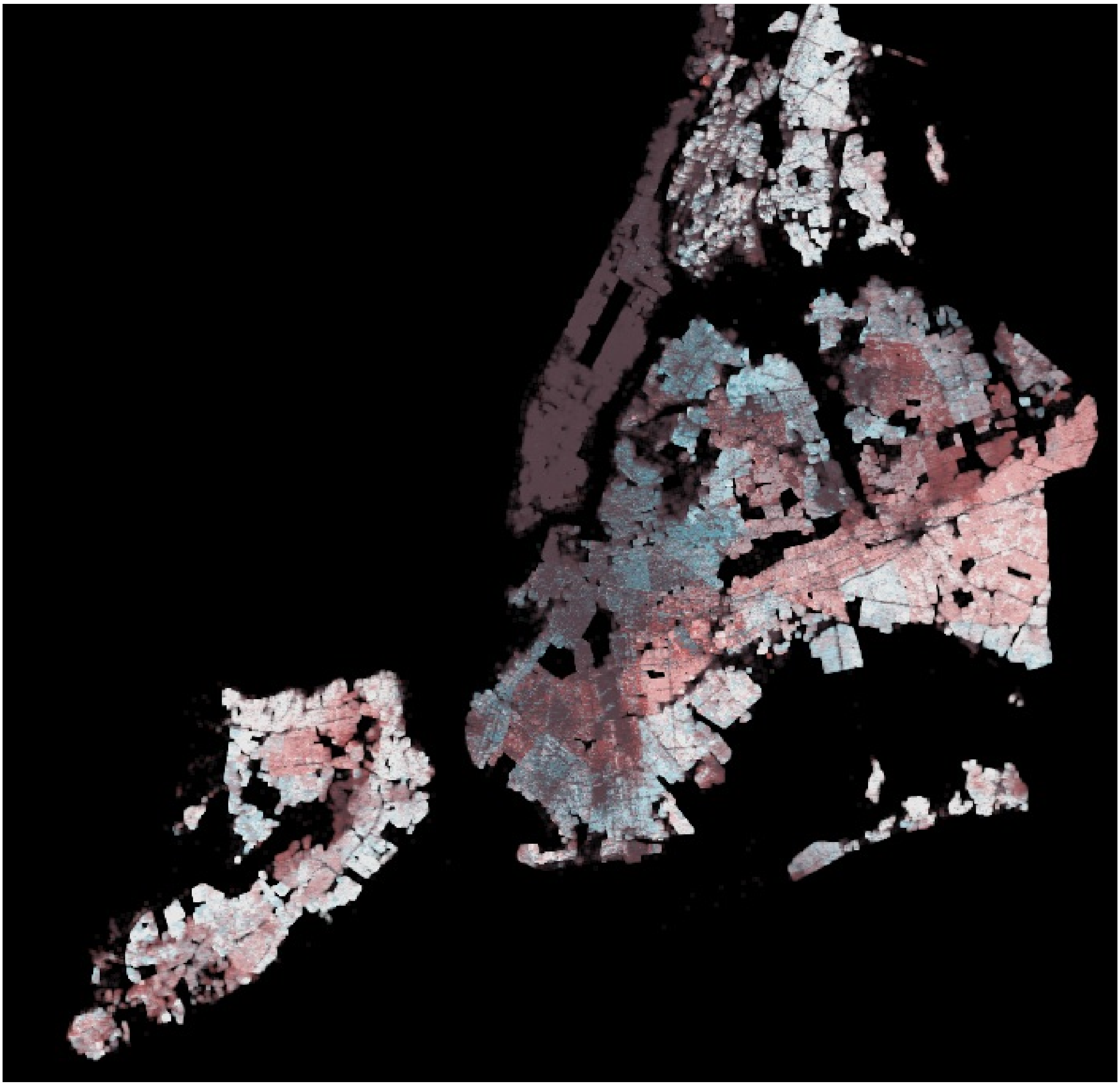
```
ny['AssessStruct'] = ny['assesstot'] - ny['assessland']

al33 = np.percentile(ny['assessland'].values, 33)
al66 = np.percentile(ny['assessland'].values, 66)
as33 = np.percentile(ny['AssessStruct'].values, 33)
as66 = np.percentile(ny['AssessStruct'].values, 66)

ny['AssessLand3tile'] = ny['assessland'].apply(lambda x: 1 if x < al33 else 2 if x < al6
6 else 3)
ny['AssessStruct3tile'] = ny['AssessStruct'].apply(lambda x: 1 if x < as33 else 2 if x <
as66 else 3)
ny['AssessAll3tile'] = ny['AssessLand3tile'].apply(str) + ny['AssessStruct3tile'].apply(
str) + 'c'
ny['AssessAll3tile'] = pd.Categorical(ny['AssessAll3tile'])

color_key = {'11c': '#e8e8e8', '12c': '#b0d5df','13c': '#64acbe','21c': '#e4acac', '22c'
: '#ad9ea5', '23c': '#627f8c',
    '31c': '#c85a5a','32c': '#985356', '33c': '#574249'}

# greener is Assessed Land, redder is Assessed Total, Maroon is both,
```

The Bivariate map will show in blue green Assesed land values, red tones will show assesed total and brown will be both

In [13]:

```
NewYorkCity   = (( -74.29,  -73.69), (40.49, 40.92))
cvs = ds.Canvas(700, 700, *NewYorkCity)
agg = cvs.points(ny, 'lon', 'lat', ds.count_cat('AssessAll3tile'))
view = tf.shade(agg, color_key=color_key)
export(tf.spread(view, px=1), 'bivariate')
```

Out[13]:



As you can see the relationship between land and construct valuation heavily locates in the manhattan area and the bivariate graph show specific some of the "odd" valuations tha give the devlopers where land has been valued "low" and has great potential for development or viceversa.