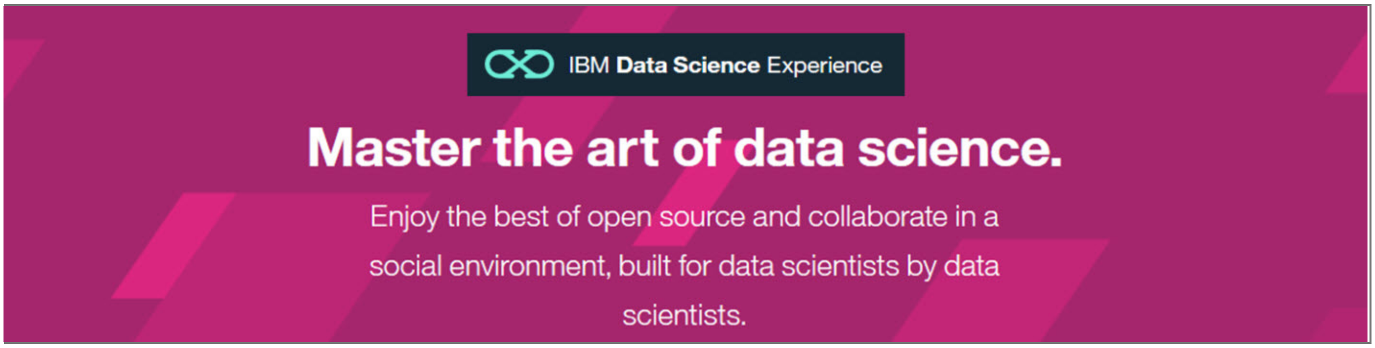


Master the art of data science.

Enjoy the best of open source and collaborate in a social environment, built for data scientists by data scientists.

Data Science Experience Tutorial

November 2016

Contents

Intro IBM Data Science Experience

IBM Data Science Experience is an interactive, collaborative, cloud-based environment where data scientists can use multiple tools to activate their insights. Data scientists can use the best of open source, tap into IBM's unique features, grow their capabilities, and share their successes. In addition to all the current features, many new capabilities are being added including the ability to ingest Object Storage data with a single click, an enhanced user interface for version control, a facility to comment or chat about a notebook with others, and many more!

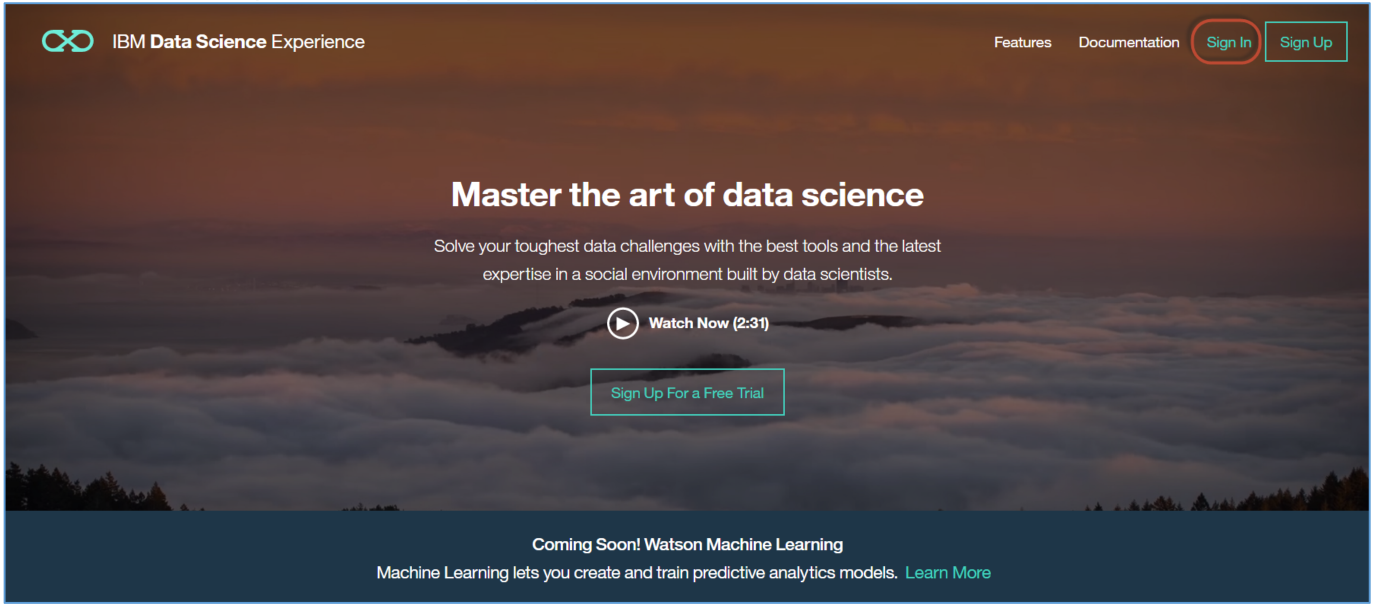
This tutorial will use real restaurant inspection records for most of the state of New York. Insights and visualizations using maps and charts will be achieved using this data.

If you would like to view a short 4 minute recorded demo of this tutorial, please go to http://ibm.biz/nyrestaurantsdemo.

Step 1 Sign In to IBM Data Science Experience and Create a Project

It is assumed that you have an 'IBM Data Science Experience' account. If not, please click 'Sign Up' or 'Sign Up for a Free Trial' at http://datascience.ibm.com.

a) Open your browser and navigate to http://datascience.ibm.com. Then select 'Sign In' and enter your IBMid or email and password when prompted.



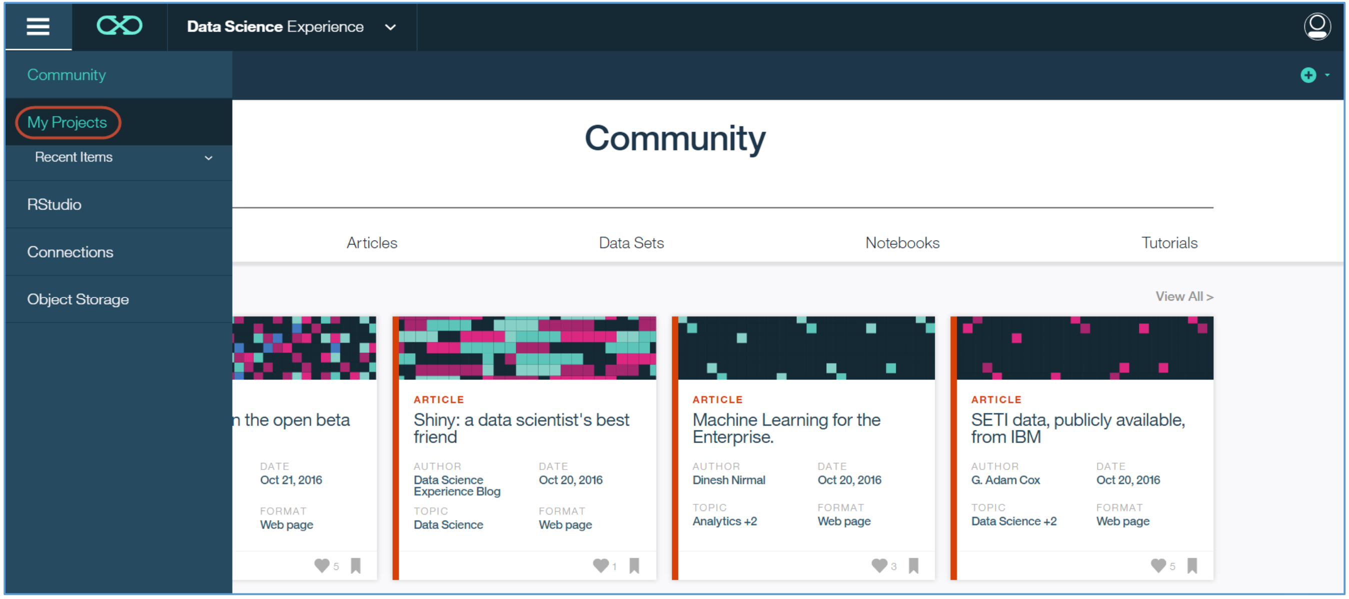
Master the art of data science

Solve your toughest data challenges with the best tools and the latest expertise in a social environment built by data scientists.

Coming Soon! Watson Machine Learning

Machine Learning lets you create and train predictive analytics models. Learn More

b) After a successful sign in, click on the ‘hamburger’ icon in the top left and select ‘My Projects’.



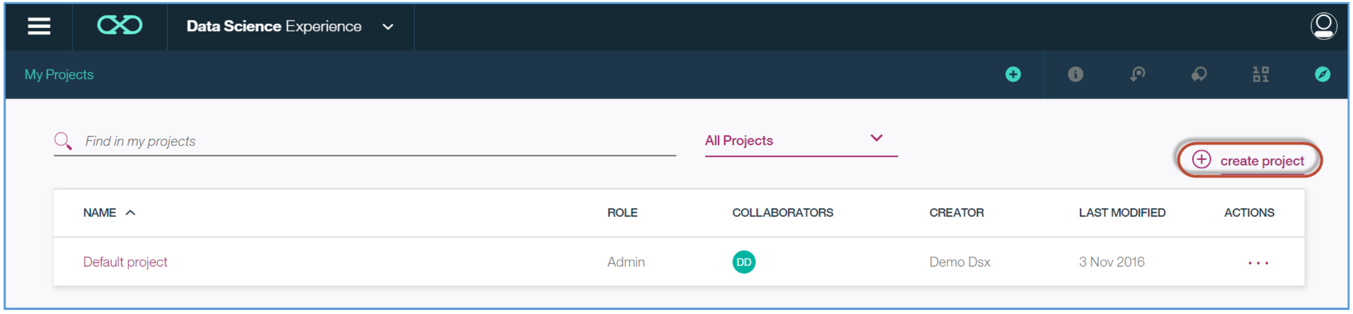
n the open beta
DATE Oct 21, 2016
FORMAT Web page

ARTICLE
Shiny: a data scientist's best friend
AUTHOR Data Science Experience Blog
DATE Oct 20, 2016
TOPIC Data Science
FORMAT Web page

ARTICLE
Machine Learning for the Enterprise.
AUTHOR Dinesh Nirmal
DATE Oct 20, 2016
TOPIC Analytics +2
FORMAT Web page

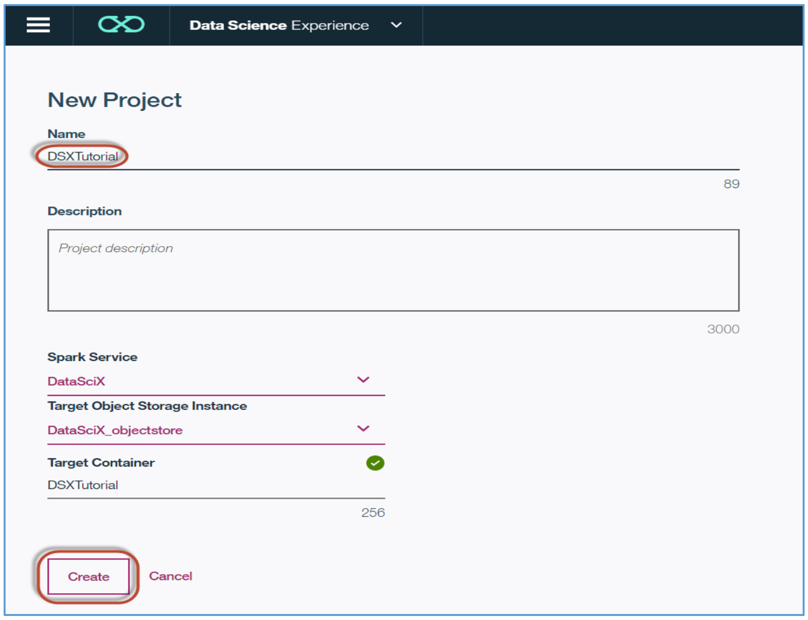
ARTICLE
SETI data, publicly available, from IBM
AUTHOR G. Adam Cox
DATE Oct 20, 2016
TOPIC Data Science +2
FORMAT Web page

c) Click on ‘create project’.



NAME	ROLE	COLLABORATORS	CREATOR	LAST MODIFIED	ACTIONS
Default project	Admin	DD	Demo Dsx	3 Nov 2016	...

d) Then enter ‘DSXTutorial’ for ‘Name’. This will also fill in ‘Target Container’ for Object Storage. Now click on ‘Create’.



New Project

Name
DSXTutorial
89

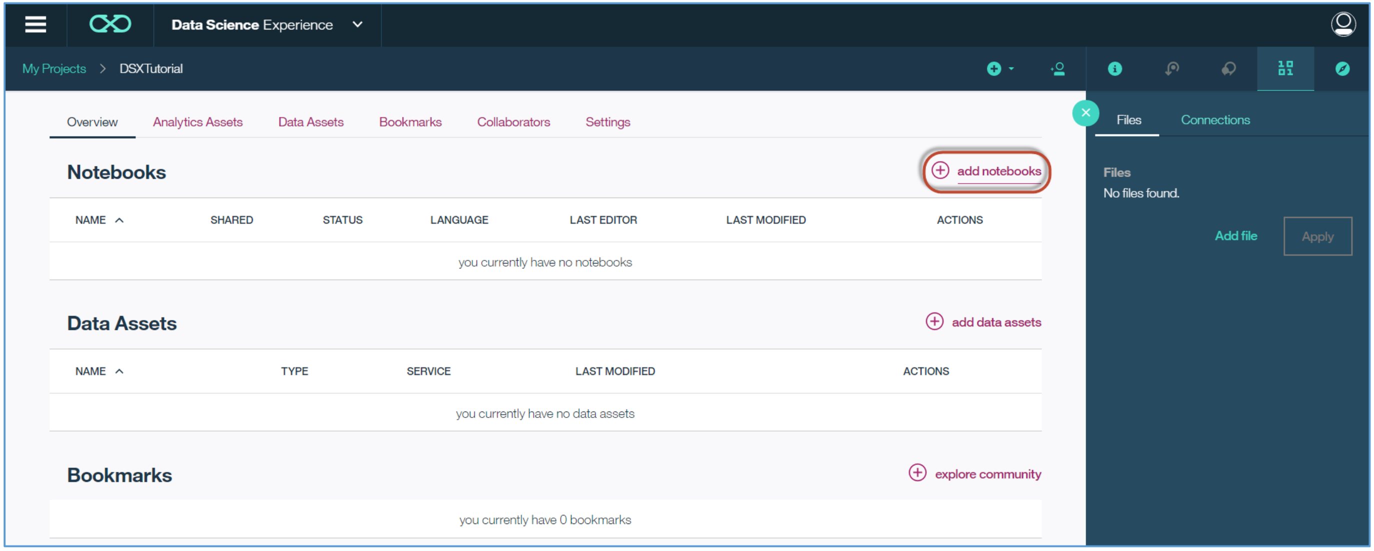
Description
Project description
3000

Spark Service
DataSciX

Target Object Storage Instance
DataSciX_objectstore

Target Container
DSXTutorial
256

e) The ‘DSXTutorial’ project page organizes notebooks, data assets and bookmarks associated with this project.



Notebooks

NAME	SHARED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS

you currently have no notebooks

Data Assets

NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS

you currently have no data assets

Bookmarks

you currently have 0 bookmarks

Files

No files found.

Step 2 Create and Work with a Notebook

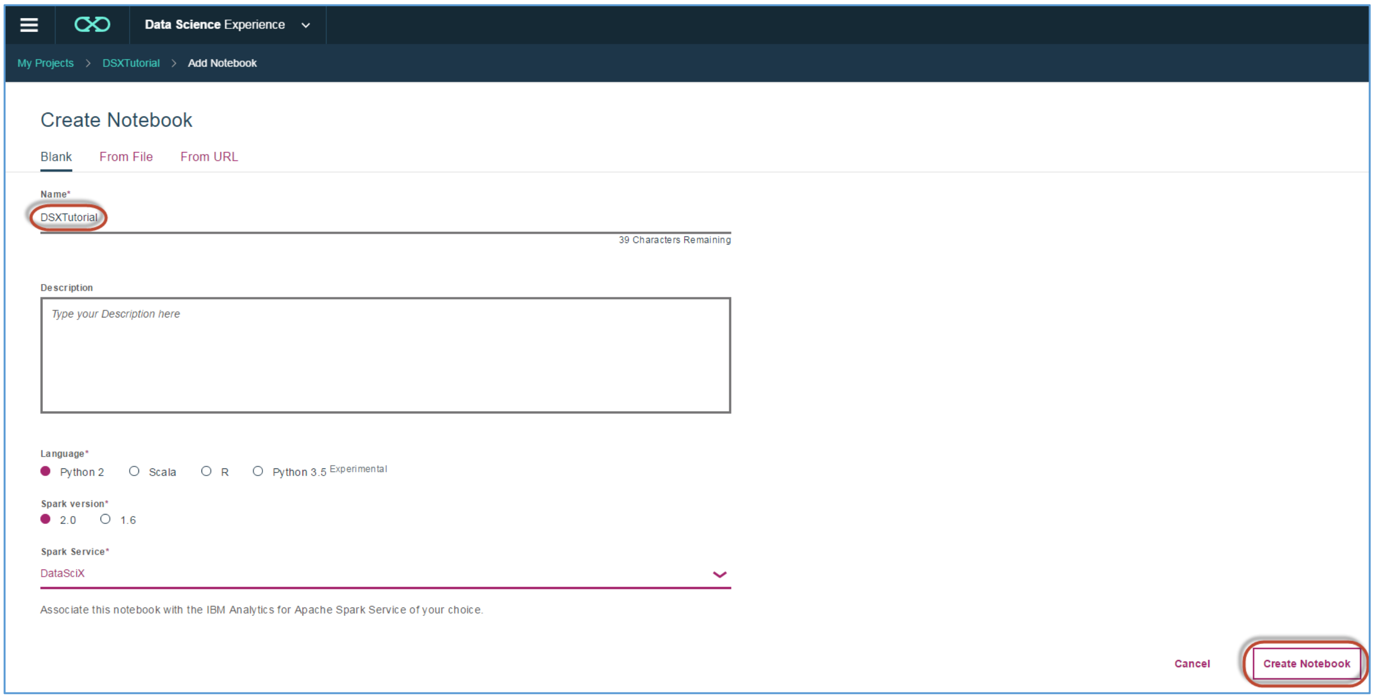
The following will walk through the steps to gain insights and visualization from this data.

Alternatively, you can view this shared notebook from Data Science Experience by going to http://ibm.biz/nyrestaurantsdsx.

As an additional alternative, a completed notebook can imported into your account by clicking on ‘add notebooks’ and using the ‘From URL’ option in ‘Create Notebook’ using the following URL. http://ibm.biz/nyrestaurantsnotebook.

To proceed by stepping through the process, please continue with the following.

f) To create a notebook, click on the ‘add notebooks’ emphasized in the previous screen capture. On the next page, enter ‘DSXTutorial’ for the ‘Name’, select ‘Python 2’ for the ‘Language’, and ‘2.0’ for the ‘Spark Version’. Then click ‘Create Notebook’.



Create Notebook

Name*: DSXTutorial

39 Characters Remaining

Description: Type your Description here

Language*: ● Python 2 ○ Scala ○ R ○ Python 3.5 Experimental

Spark version*: ● 2.0 ○ 1.6

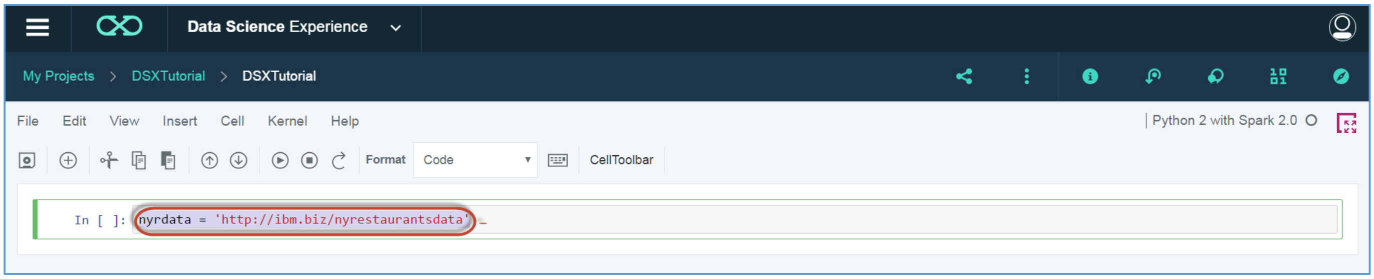
Spark Service*: DataSciX

Associate this notebook with the IBM Analytics for Apache Spark Service of your choice.

g) This notebook will provide insights from official restaurant inspection records for most of New York State and provide visualizations of that data. This data is available at New York State Food Service Establishment: Last Inspection. A raw extract was taken in October of 2016 and is located at http://ibm.biz/nyrestaurantsdata.

Please enter the following into the code cell, then execute the code by clicking on the 'play' icon or using 'Shift-Enter'

```
nyrdata = 'http://ibm.biz/nyrestaurantsdata'
```

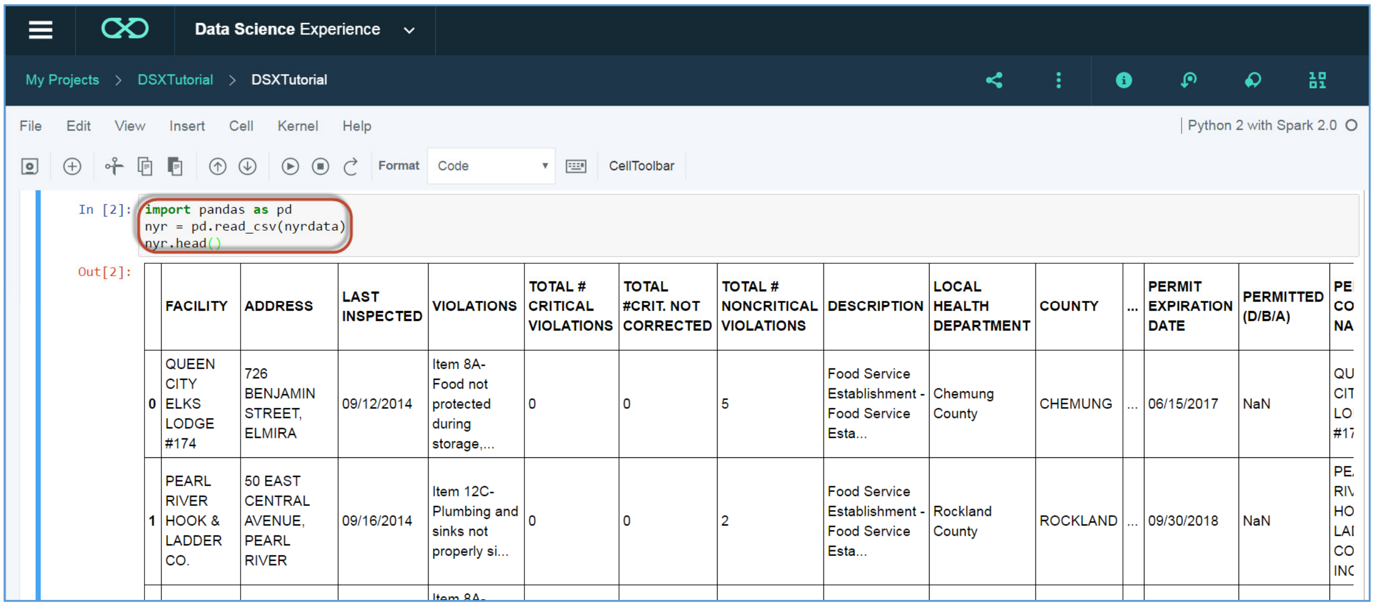


```
In [ ]: nyrdata = 'http://ibm.biz/nyrestaurantsdata'
```

h) Now, the csv (comma separated values) data will be read into a Pandas dataframe (nyr) and the first 5 records will be displayed using the 'head()' method.

Please enter the following into the next code cell and execute the code.

```
import pandas as pd
nyr = pd.read_csv(nyrdata)
nyr.head()
```


Data Science Experience
My Projects > DSXTutorial > DSXTutorial
File Edit View Insert Cell Kernel Help
Python 2 with Spark 2.0
Format
Code
CellToolbar

```
In [2]: import pandas as pd
        nyr = pd.read_csv(nyrdata)
        nyr.head()
```

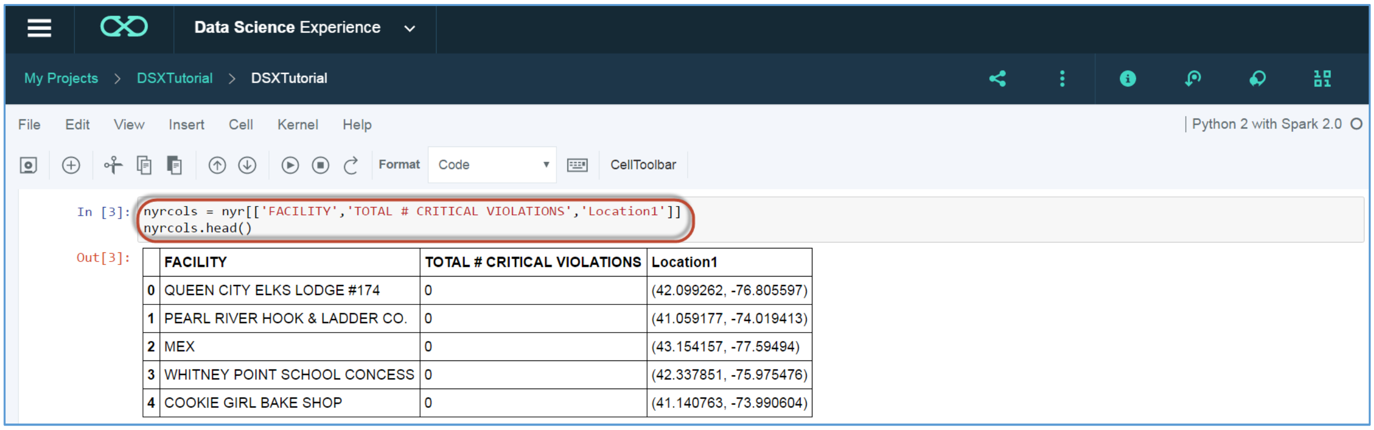
Out[2]:

	FACILITY	ADDRESS	LAST INSPECTED	VIOLATIONS	TOTAL # CRITICAL VIOLATIONS	TOTAL #CRIT. NOT CORRECTED	TOTAL # NONCRITICAL VIOLATIONS	DESCRIPTION	LOCAL HEALTH DEPARTMENT	COUNTY	...	PERMIT EXPIRATION DATE	PERMITTED (D/B/A)	PE CO NA
0	QUEEN CITY ELKS LODGE #174	726 BENJAMIN STREET, ELMIRA	09/12/2014	Item 8A-Food not protected during storage,...	0	0	5	Food Service Establishment - Food Service Esta...	Chemung County	CHEMUNG	...	06/15/2017	NaN	QU CIT LO #17
1	PEARL RIVER HOOK & LADDER CO.	50 EAST CENTRAL AVENUE, PEARL RIVER	09/16/2014	Item 12C-Plumbing and sinks not properly si...	0	0	2	Food Service Establishment - Food Service Esta...	Rockland County	ROCKLAND	...	09/30/2018	NaN	PE. RI\ HO LAI CO INC
				Item 8A-										

The data has been ingested and displayed in an easy-to-read table. As you can see, data can be accessed by using one line of code. Data can be ingested from Cloudant, DashDB, Object Storage, relational databases, and many others.

i) Another dataframe will be created that will only contain the columns that are pertinent to this analysis. The 'head()' method will display the first 5 records of this dataframe.

Please enter the following into the next code cell and execute the code
nyrcols = nyr[['FACILITY','TOTAL # CRITICAL VIOLATIONS','Location1']]
nyrcols.head()


Data Science Experience
My Projects > DSXTutorial > DSXTutorial
File Edit View Insert Cell Kernel Help
Python 2 with Spark 2.0
Format
Code
CellToolbar

```
In [3]: nyrcols = nyr[['FACILITY','TOTAL # CRITICAL VIOLATIONS','Location1']]
        nyrcols.head()
```

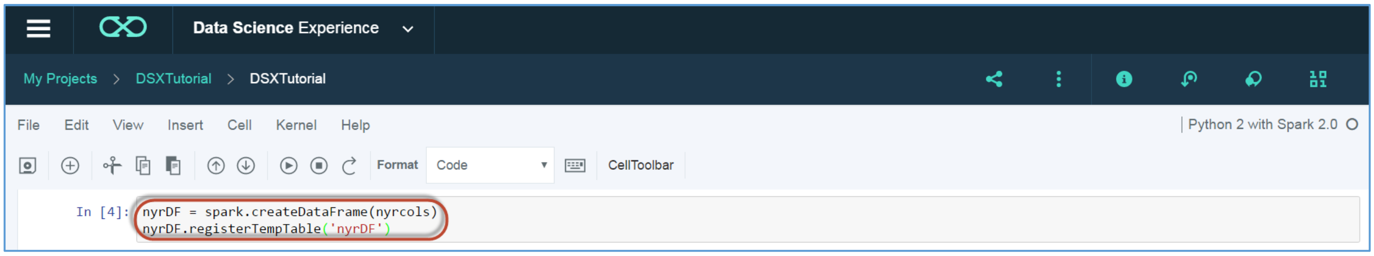
Out[3]:

	FACILITY	TOTAL # CRITICAL VIOLATIONS	Location1
0	QUEEN CITY ELKS LODGE #174	0	(42.099262, -76.805597)
1	PEARL RIVER HOOK & LADDER CO.	0	(41.059177, -74.019413)
2	MEX	0	(43.154157, -77.59494)
3	WHITNEY POINT SCHOOL CONCESS	0	(42.337851, -75.975476)
4	COOKIE GIRL BAKE SHOP	0	(41.140763, -73.990604)

j) At this point, the data will be transformed into a Spark dataframe 'nyrDF' and a table will be registered. Spark dataframes are conceptually equivalent to a table in a relational database or a dataframe in R/Python, but with richer optimizations under the hood. A table that is registered can be used in subsequent SQL statements.

Please enter the following into the next code cell and execute the code.

```
nyrDF = spark.createDataFrame(nyrcols)
nyrDF.registerTempTable('nyrDF')
```

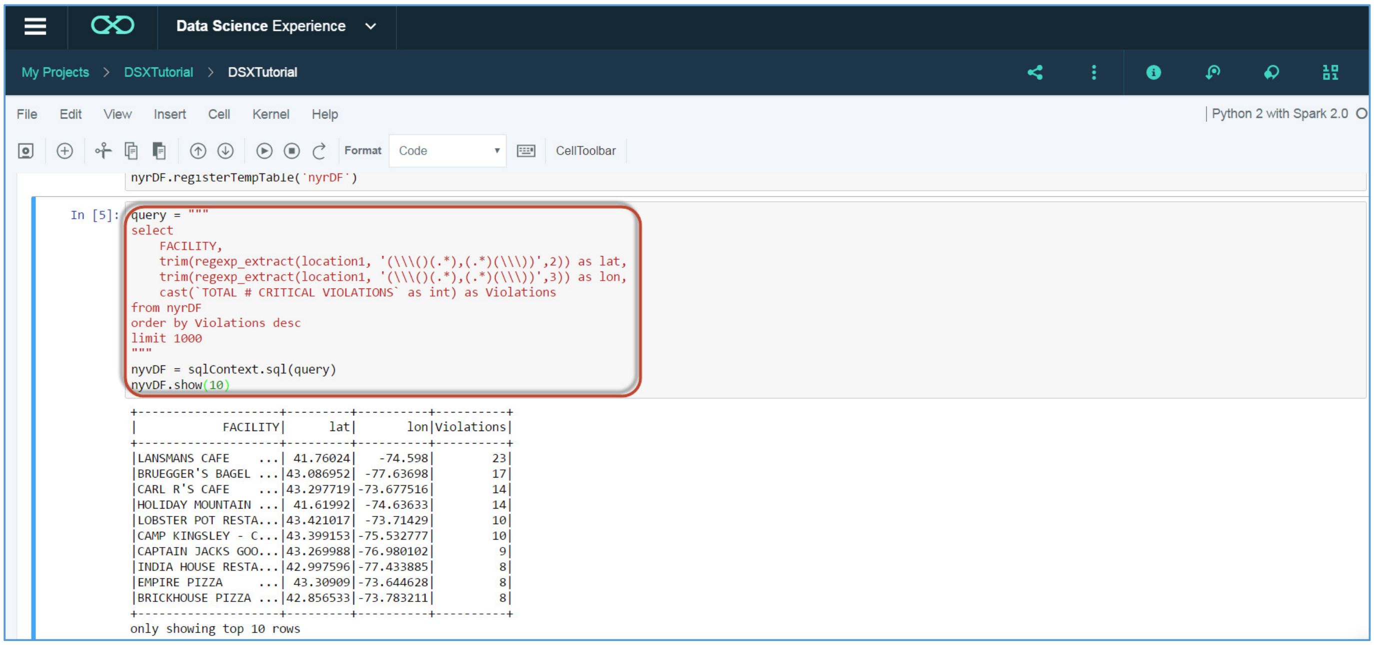

Data Science Experience
My Projects > DSXTutorial > DSXTutorial
File Edit View Insert Cell Kernel Help
Python 2 with Spark 2.0
Format
Code
CellToolbar

```
In [4]: nyrDF = spark.createDataFrame(nyrcols)
        nyrDF.registerTempTable('nyrDF')
```

k) A Spark dataframe 'nyvDF' will be created using SQL that will contain the restaurant name (FACILITY), latitude, longitude and violations. Note that the latitude and longitude are combined in the final column (Location1) of the retrieved data. They will be extracted separately using regular expressions in the SQL. The results are ordered by number of violations in descending order and the top 10 are displayed.

Please enter the following into the next code cell and execute the code.

```
query = """
select
   FACILITY,
   trim(regexp_extract(location1, '(\\\()(.*),(.*)(\\\))',2)) as lat,
   trim(regexp_extract(location1, '(\\\()(.*),(.*)(\\\))',3)) as lon,
   cast(`TOTAL # CRITICAL VIOLATIONS` as int) as Violations
from nyrDF
order by Violations desc
limit 1000
"""
nyvDF = sqlContext.sql(query)
nyvDF.show(10)
```


Data Science Experience
My Projects > DSXTutorial > DSXTutorial
File Edit View Insert Cell Kernel Help
Python 2 with Spark 2.0
Format Code CellToolbar

```
nyrDF.registerTempTable('nyrDF')
```

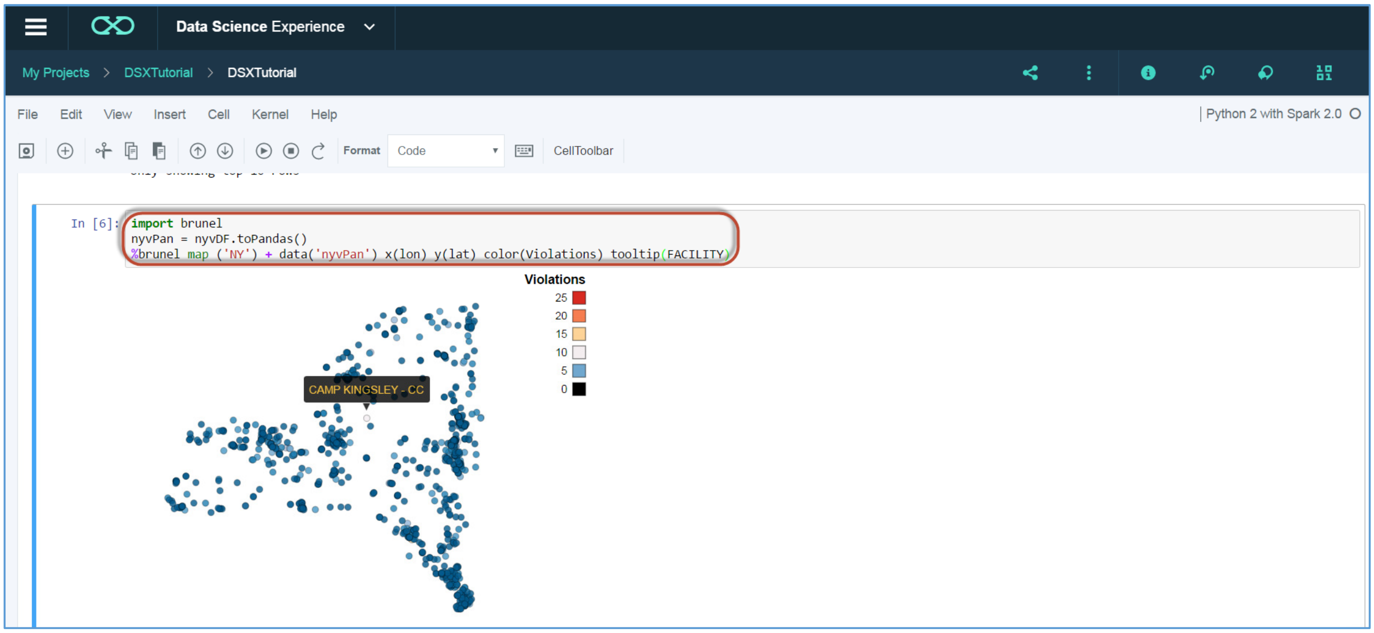
```
In [5]: query = """
select
    FACILITY,
    trim(regexp_extract(location1, '(\\\()(.*),(.*)(\\\))',2)) as lat,
    trim(regexp_extract(location1, '(\\\()(.*),(.*)(\\\))',3)) as lon,
    cast(`TOTAL # CRITICAL VIOLATIONS` as int) as Violations
from nyrDF
order by Violations desc
limit 1000
"""

nyvDF = sqlContext.sql(query)
nyvDF.show(10)
```

```
+--------------------+---------+----------+----------+
|            FACILITY|      lat|       lon|Violations|
+--------------------+---------+----------+----------+
|LANSMANS CAFE    ...| 41.76024|   -74.598|        23|
|BRUEGGER'S BAGEL ...|43.086952| -77.63698|        17|
|CARL R'S CAFE    ...|43.297719|-73.677516|        14|
|HOLIDAY MOUNTAIN ...| 41.61992| -74.63633|        14|
|LOBSTER POT RESTA...|43.421017| -73.71429|        10|
|CAMP KINGSLEY - C...|43.399153|-75.532777|        10|
|CAPTAIN JACKS GOO...|43.269988|-76.980102|         9|
|INDIA HOUSE RESTA...|42.997596|-77.433885|         8|
|EMPIRE PIZZA     ...| 43.30909|-73.644628|         8|
|BRICKHOUSE PIZZA ...|42.856533|-73.783211|         8|
+--------------------+---------+----------+----------+
only showing top 10 rows
```

l) Brunel visualization will be used to map the latitude and longitude to a New York state map. Colors represent the number of violations as noted in the key.

Please enter the following code into the next code cell and execute the code.
import brunel
nyvPan = nyvDF.toPandas()
%brunel map ('NY') + data('nyvPan') x(lon) y(lat) color(Violations) tooltip(FACILITY)



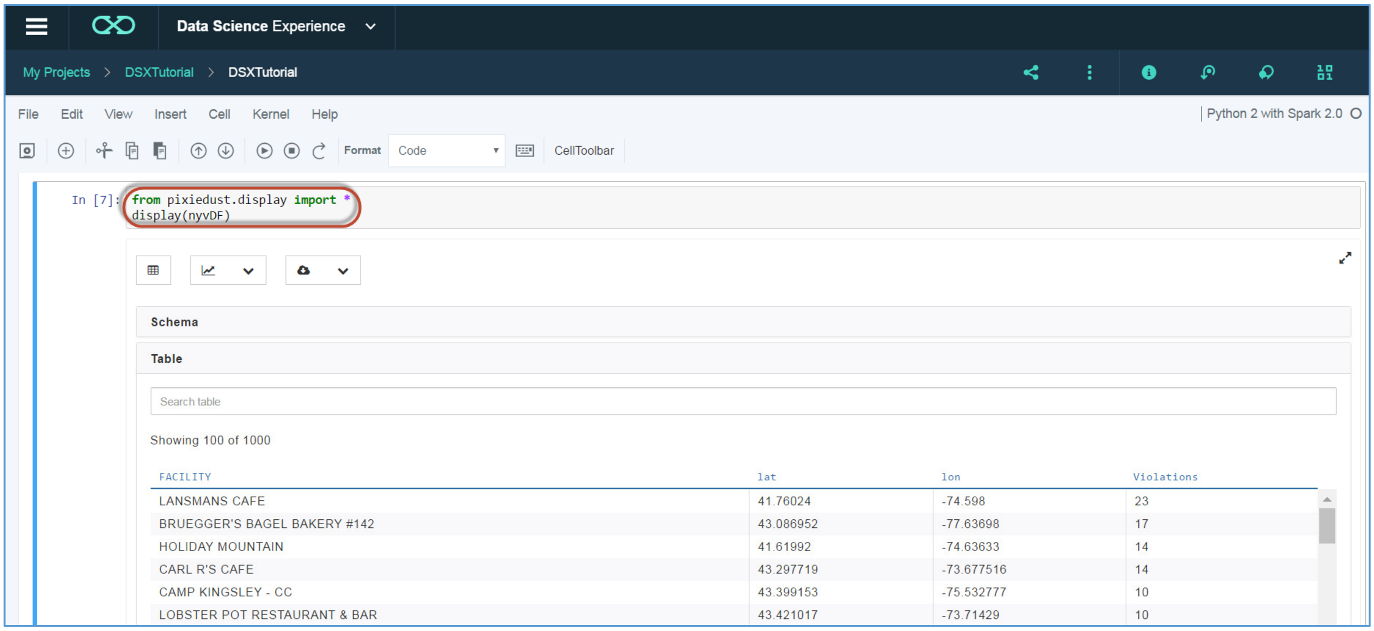
```
In [6]: import brunel
nyvPan = nyvDF.toPandas()
%brunel map ('NY') + data('nyvPan') x(lon) y(lat) color(Violations) tooltip(FACILITY)
```

m) One of the many key strengths of Data Science Experience is the ability to easily search and quickly learn about various topics. For example, to find articles, tutorials or notebooks on Brunel, click on the 'link' icon on the top right hand corner of this web page ('Find Resources in the Commuity'). A side palette will appear where you can enter 'Brunel' or other topics of interest. Related articles, tutorials, notebooks, data cards will be displayed.

n) Pixiedust provides charting and visualization. It is an open source Python library that works as an add-on to Jupyter notebooks to improve the user experience of working with data

Please enter the following code in the next code cell and execute the code.

```
from pixiedust.display import *
display(nyvDF)
```



In [7]:
```
from pixiedust.display import *
display(nyvDF)
```

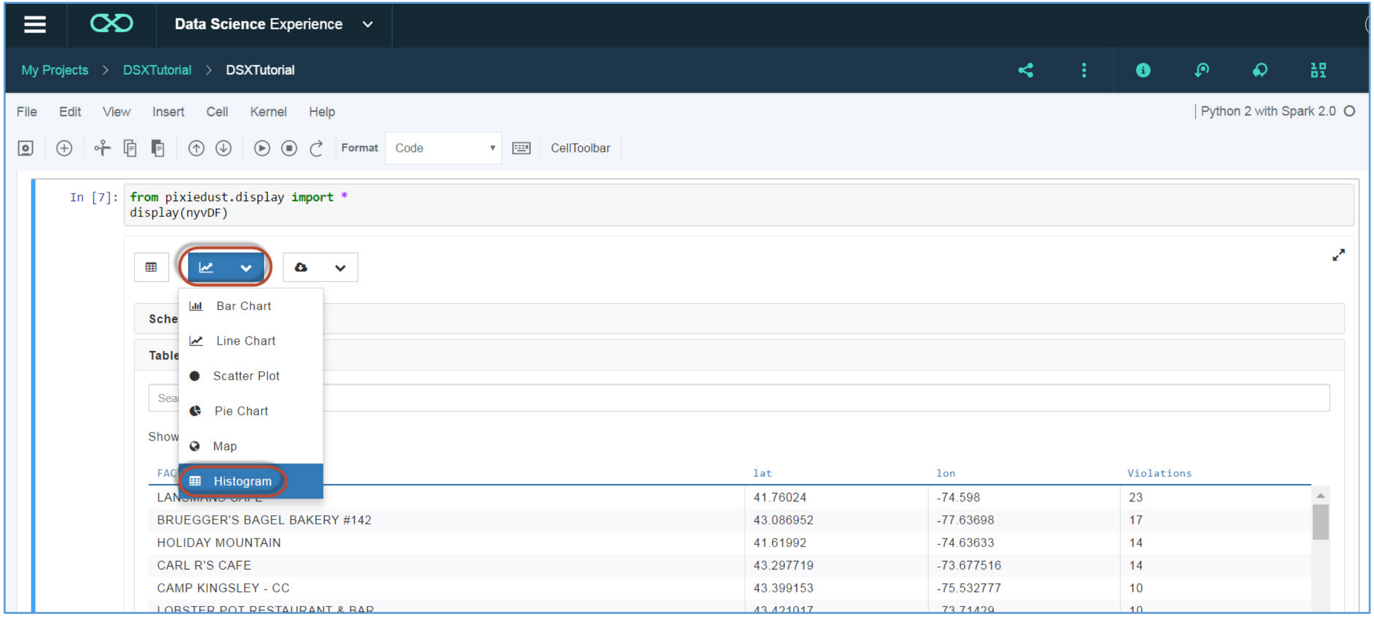
Search table

Showing 100 of 1000

FACILITY	lat	lon	Violations
LANSMANS CAFE	41.76024	-74.598	23
BRUEGGER'S BAGEL BAKERY #142	43.086952	-77.63698	17
HOLIDAY MOUNTAIN	41.61992	-74.63633	14
CARL R'S CAFE	43.297719	-73.677516	14
CAMP KINGSLEY - CC	43.399153	-75.532777	10
LOBSTER POT RESTAURANT & BAR	43.421017	-73.71429	10

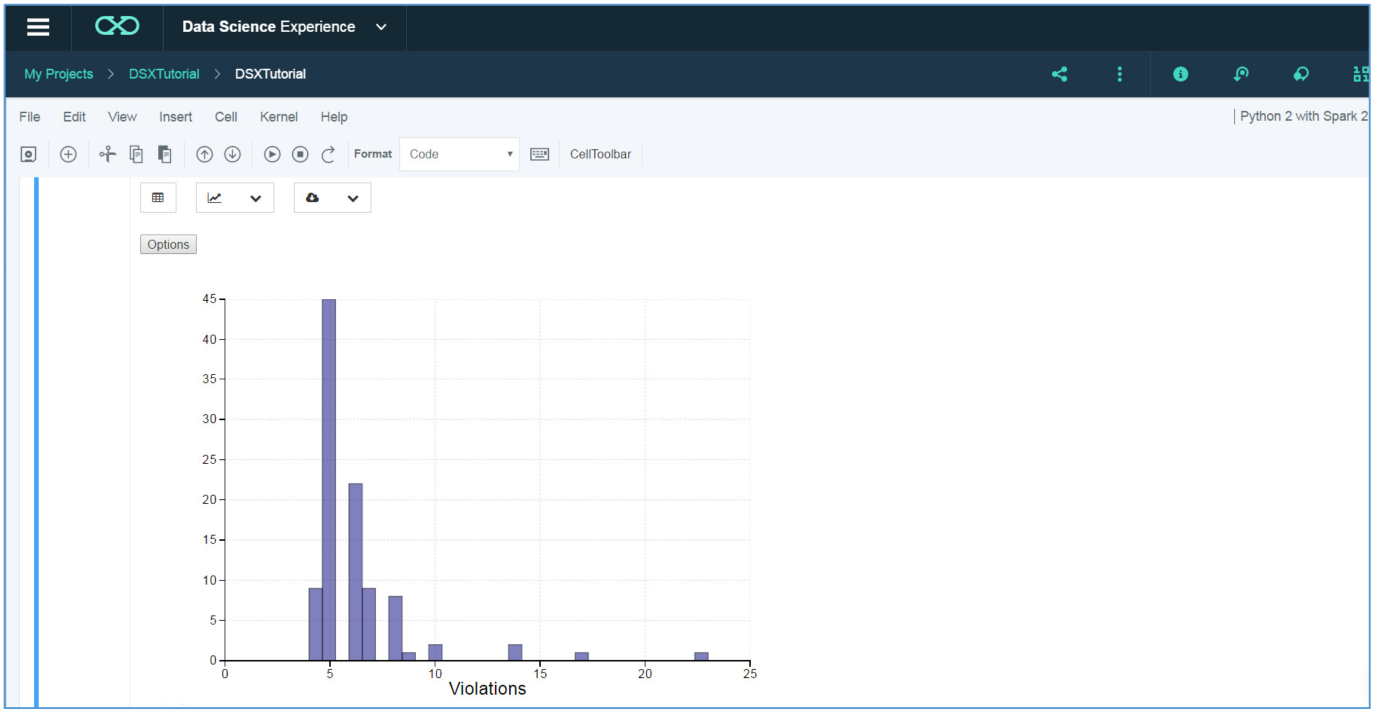
o) If you hover over the lonely lighter colored dot in the middle of the New York State map, you can see that it is for 'CAMP KINGSLEY - CC'. By starting to type the value 'camp' in the 'Search table' text field above, the record will be displayed. In addition, the data can be downloaded as a file, or stashed to Cloudant or Object Storage

p) In addition, many charting options are available. For example, please look at the histogram.



```
In [7]: from pixiedust.display import *
        display(nyvDF)
```

FAC	lat	lon	Violations
LANSMANS CAFE	41.76024	-74.598	23
BRUEGGER'S BAGEL BAKERY #142	43.086952	-77.63698	17
HOLIDAY MOUNTAIN	41.61992	-74.63633	14
CARL R'S CAFE	43.297719	-73.677516	14
CAMP KINGSLEY - CC	43.399153	-75.532777	10
LOBSTER POT RESTAURANT & BAR	43.421017	-73.71429	10



q) In just a few notebook cells, data was ingested, manipulated, visualized and yielded insights. Much more capability, including machine learning, could be leveraged with IBM Data Science Experience. This is just the tip of the iceberg!

Congratulations, you have completed this exercise.

Great work and congratulations, you have completed this exercise.

Screen captures in this document may vary slightly from yours.

NOTES

The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way.