Extra assignment

Topic:

Use Austin Restaurant inspection report data and spark to get answer to critical consumer questions such as which area rest has more violations in past year etc...

City inspectors found multiple instances of the most serious type of health and sanitary code violations at Austin's restaurants and food service locations, according to a Globe review of municipal data

Analytics MyNotebook272 and Bluemix Apache Spark

My approach:

Data set

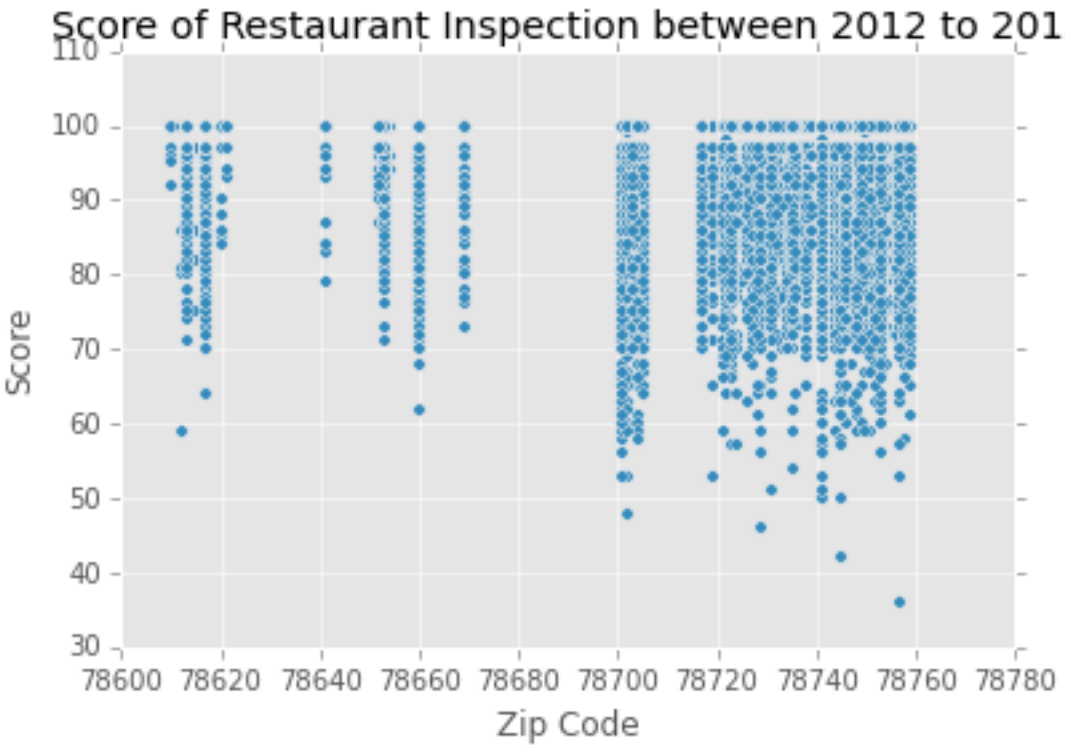
I will use **Apache Spark** to analyze raw Restaurant Inspection Scores data from data.austintexas.gov. More specifically we use the raw restaurant scores for inspections performed within the last three years. Online search of this data set also available at: http://www.ci.austin.tx.us/health/restaurant/search.cfm
The format of the raw data in the base file consists of the following:

Restaurant Name	Zip Code	Inspection Date	Score	Address	Facility ID	Process Description
15th Street Cafe	78701	12/5/13	97	"303 W 15TH ST AUSTIN, TX 78701 (30.277501963873, -97.7425547470704)"	2801033	Routine Inspection
15th Street Cafe	78701	12/2/14	93	"303 W 15TH ST AUSTIN, TX 78701 (30.277501963873, -97.7425547470704)"	2801033	Routine Inspection
15th Street Cafe	78701	6/18/14	97	"303 W 15TH ST AUSTIN, TX 78701 (30.277501963873, -97.7425547470704)"	2801033	Routine Inspection
...	...	...	...	...	...	...

where columns contain a Restaurant Name, Zip Code, Inspection Date, Score, Address, Facility ID, Process Description.

Data Analysis:

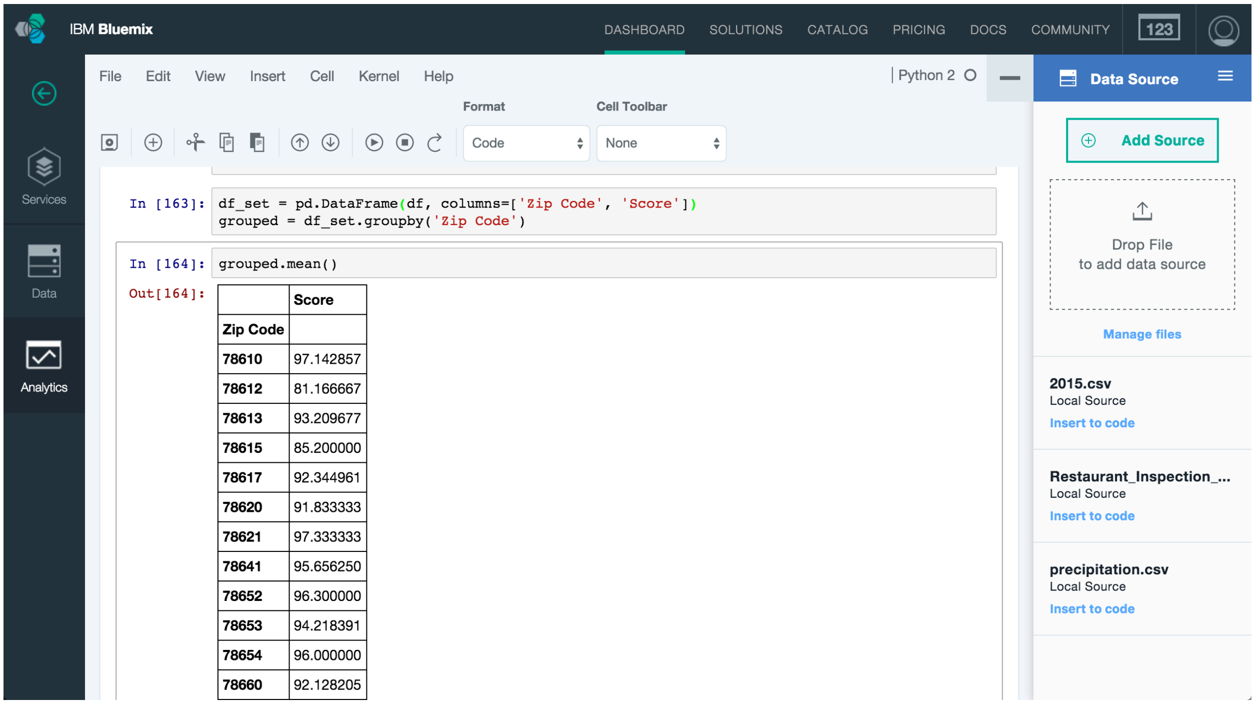
I look deep into the dataset, and find use the restaurant zip code can cluster the inspection score like blow:



Raise Questions

I explored and prepared this raw data. Now, I can raise more complex questions and make use of the data to answer them.

Question: What is the average inspection score of specific area (Zip Code) of Austin?



```
In [163]: df_set = pd.DataFrame(df, columns=['Zip Code', 'Score'])
          grouped = df_set.groupby('Zip Code')
```

```
In [164]: grouped.mean()
Out[164]:
```

	Score
Zip Code	
78610	97.142857
78612	81.166667
78613	93.209677
78615	85.200000
78617	92.344961
78620	91.833333
78621	97.333333
78641	95.656250
78652	96.300000
78653	94.218391
78654	96.000000
78660	92.128205

Reference:

1.Jupyter Notebook https://jupyter.org/index.html
2.IBM Blumix Tutorial https://www.ng.bluemix.net/docs/services/AnalyticsforApacheSpark/index.html