**K-Means Clustering**

**Ask Business Question**

Can the K-means clustering algorithm accurately distinguish (cluster) between the various neighborhoods of New York City (NYC):

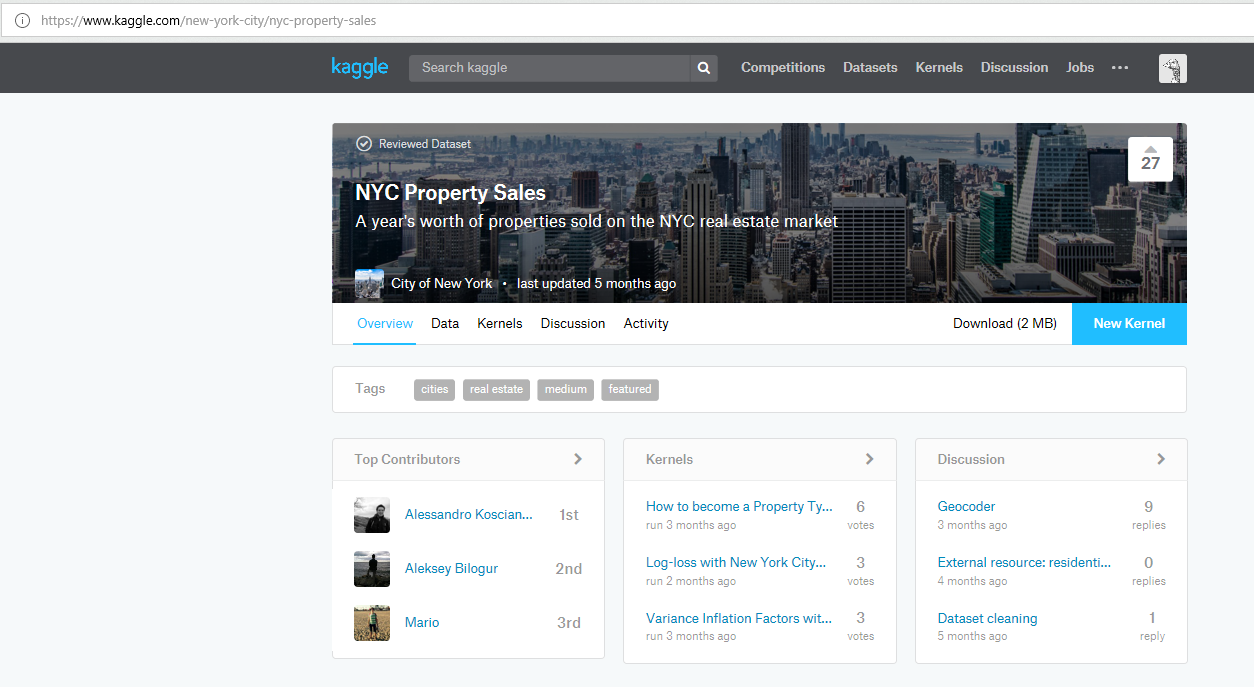
* Manhattan
* Brooklyn
* Bronx
* Queens
* Staten Island

Given a year’s worth of property sale data in the NYC real-estate market.

**0. Acquire Data**

Download and Observe the Data. The data is found at:

<https://www.kaggle.com/new-york-city/nyc-property-sales>



**1. Preprocess Data**

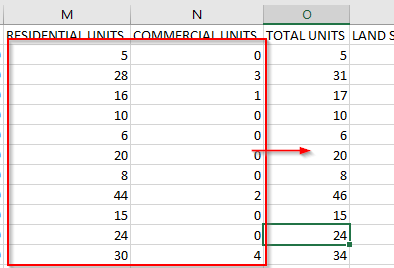
After downloading the data, it is always good to open the data and glance through what the data contains.

The K-means algorithm considers the *distance* between two vectors. Consider the following three properties:

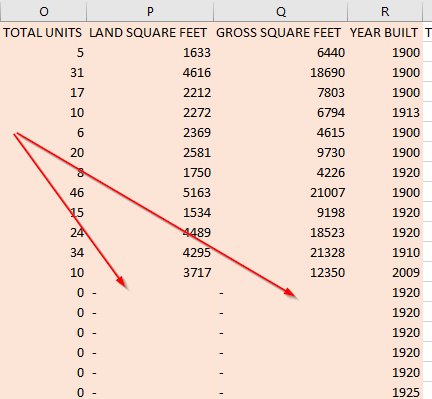
|  |  |  |
| --- | --- | --- |
| **Property** | **Sales Price** | **Zip Code** |
| A | $285,000 | 10009 |
| B | $300,000 | 10036 |
| C | $500,000 | 10011 |

We can say with confidence that properties A and B are closer in sales price than they are to property C, simply judging from the numerical differences in the sales price. However, if we consider zip codes as a numerical category (rather than symbolic), property A and C are closer. While there is some logic behind the numbering system of NYC zip codes, for the purposes of this exercise we remove this feature in our K-means exercise.

Looking at the data, we also notice some correlation between features. For example, we see that the total number of units sold is equal to the sum of the residential and commercial units sold. In this case, we simply take the aggregate ‘TOTAL UNITS’ only.



Finally, looking at the data we see in some entries the features we want are missing:



These missing values can be dealt with in multiple ways (filled with the mean of all other data or simply removed if there is a sufficient number of filled, non-missing data). In this exercise we remove these rows.

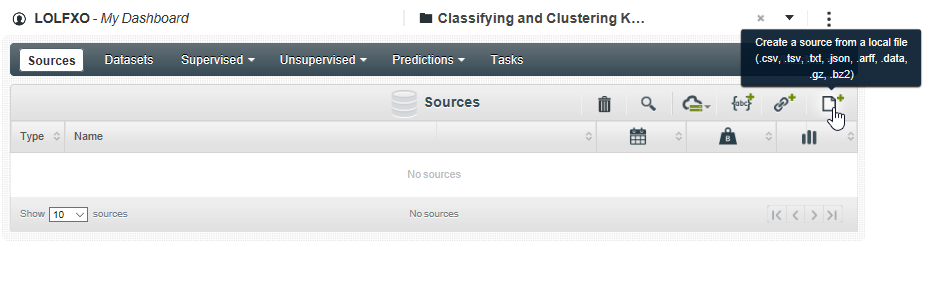
Once the features have been selected and the data cleaned (sales with any selected features missing removed), we have the following features:

* Total Units
* Land per square feet
* Gross square feet
* Year built
* Sales price
* Sales date

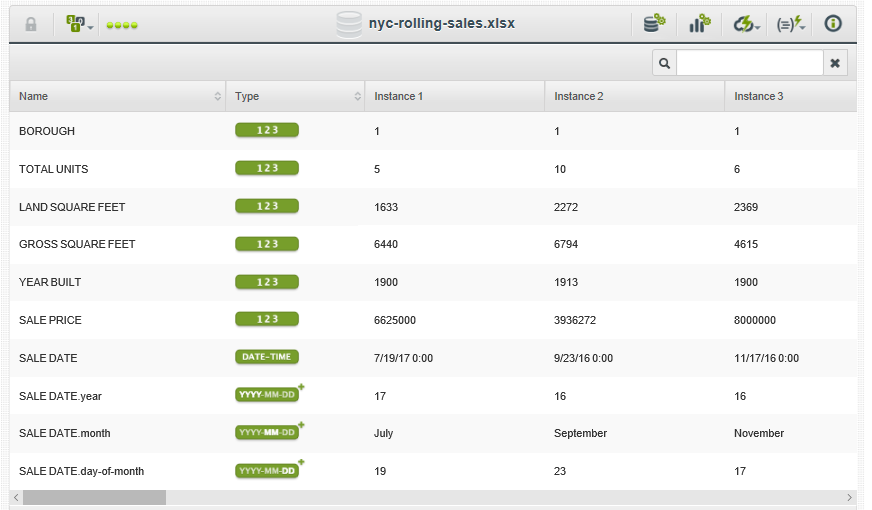
The preprocessing can take place either within the Excel course itself, or in BigML.

**2. Upload Data to BigML**

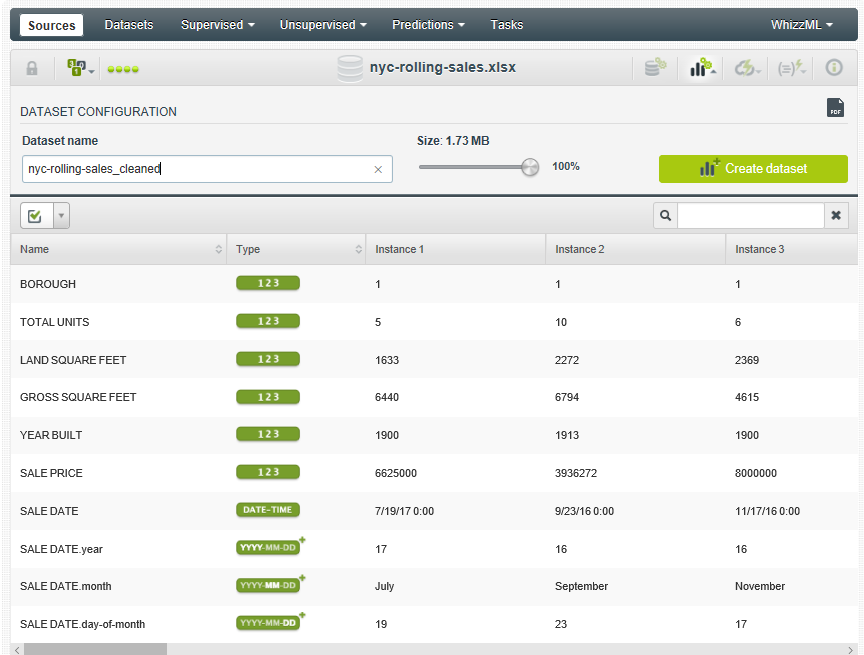
Once we know which features to consider, we upload our data BigML as a source:



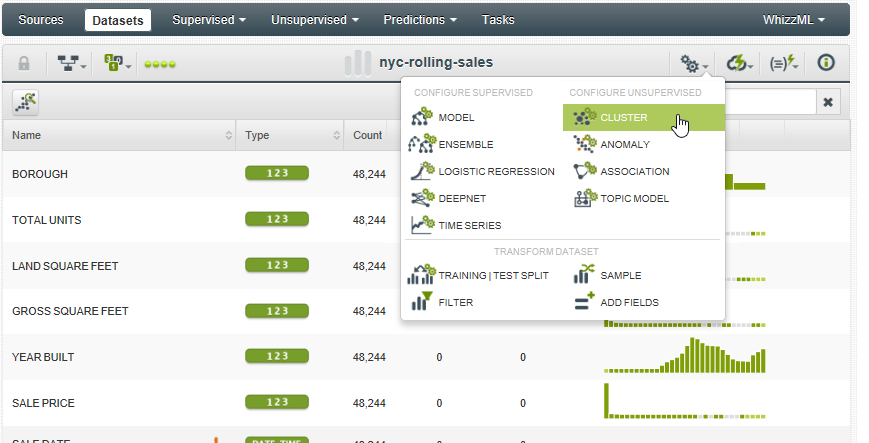
Once we have uploaded the cleaned file, clicking on the file itself we see:



Once we are happy with the data we can create a dataset:

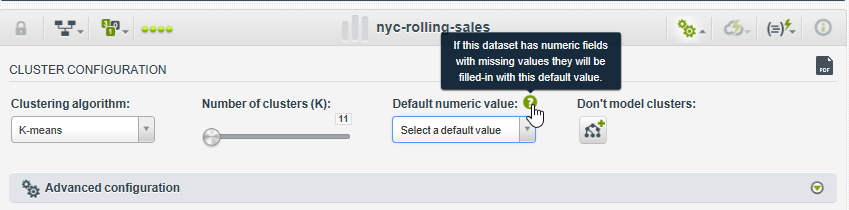


Once a dataset has been created, we can carry out our cluster analysis:

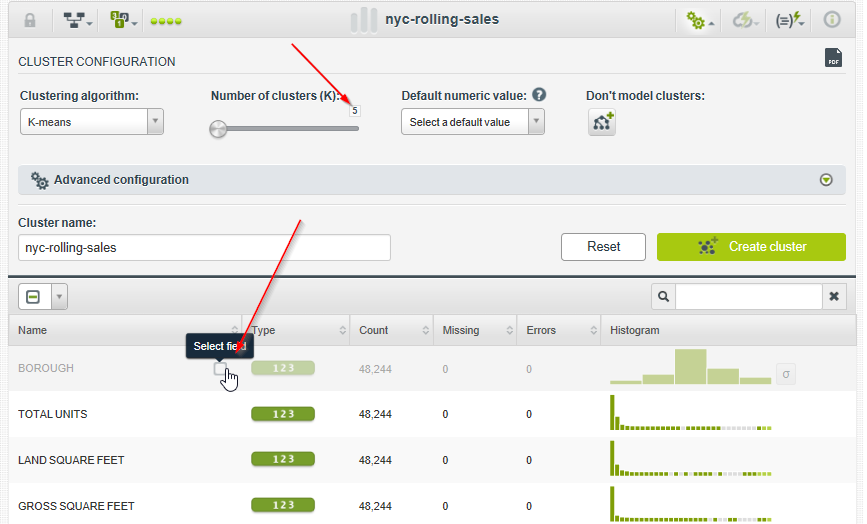


**3. K-Means Cluster**

When we come to cluster our data, we can deal with missing values to be filled with a number of different options (e.g. the mean of present data).

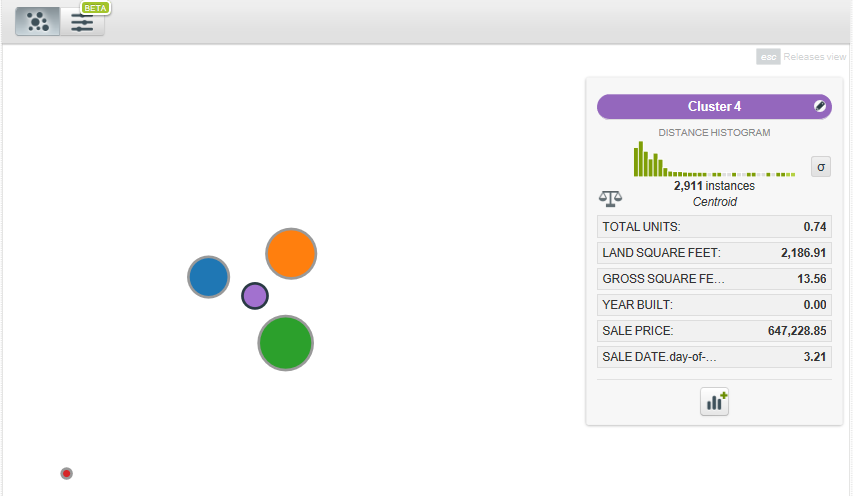


This screen also allows us to define the number of clusters. We want to set the number of clusters to be 5 (for the five boroughs), and since K-means is an unsupervised clustering algorithm, we de-select the ‘BOROUGH’ feature:



Since our data set includes sales only within 2017, we can also remove the sale date as a feature that has an input to how we cluster the property sales.

We are now ready to ‘Create cluster’. The results of the cluster are as below:



The box on the right-hand side describes the cluster selected (Cluster 4). This cluster has within it 16144 instances, with the properties of the cluster centroid as shown.

**4. Interpret Cluster Results**

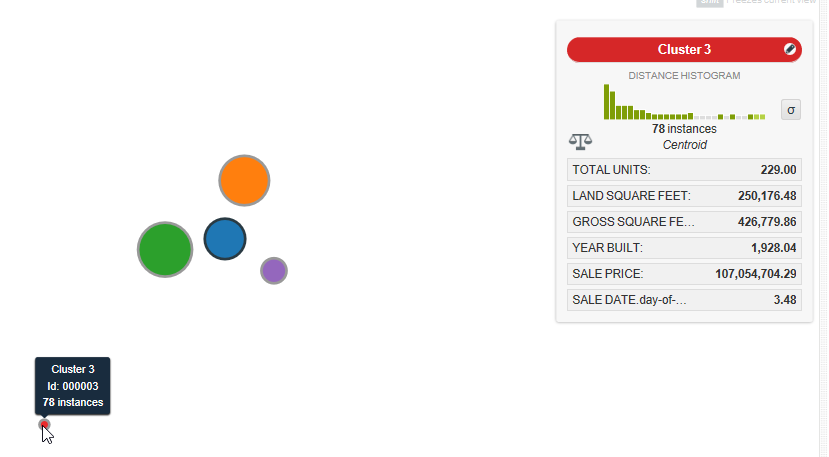


The above shows the actual statistics of each borough from the data. The cluster output from BigML is as follows:



Interestingly, BigML has clustered the data giving importance to different features. For example, the centroid of the 2911 apartments in cluster 4 has a ‘Year Built’ feature of 0 and the centroid of the 78 apartments in cluster 3 has a high total number of units and a very high sales price of $107M.

The distinction of apartments in clusters 3 with the other clusters is also shown in the cluster diagram generated by BigML:



In conclusion, while the K-means clustering algorithm, using Euclidean distance between numerical features we selected (namely Total Units, Land Square Feet, Gross Square Feet, Year Build and Sale Price) cannot distinguish between neighborhoods, the algorithm has returned an interesting segmentation of the apartments.

**5. Check Cluster Accuracy (Optional)**

The K-means clustering algorithm is unsupervised, meaning we are not meant to know the True North clustering of our data set. However, if we do have the True North clustering of the data (such as in this case, we are aware of exactly which neighborhood each property belongs to), we can assess how accurate our clustering algorithm is.