Big Data Project

Air Quality Clustering

Klim Basargin

Business Understanding

Beijing government, in order to punish and motivate companies, decided to test a dynamic taxation system.

It is required to accurately define air quality classes to assess in the future.

Goals:
Clearly separate types of emissions from industries and apply the correct tax rates.

Data Understanding

Data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017.

Data Understanding

No: row number

year: year of data in this row

month: month of data in this row

day: day of data in this row

hour: hour of data in this row

PM2.5: PM2.5 concentration (ug/m$\hat{3}$)

PM10: PM10 concentration (ug/m$\hat{3}$)

SO2: SO2 concentration (ug/m$\hat{3}$)

NO2: NO2 concentration (ug/m$\hat{3}$)

CO: CO concentration (ug/m$\hat{3}$)

O3: O3 concentration (ug/m$\hat{3}$)

TEMP: temperature (degree Celsius)

PRES: pressure (hPa)

DEWP: dew point temperature (degree Celsius)

RAIN: precipitation (mm)

wd: wind direction

WSPM: wind speed (m/s)

station: name of the air-quality monitoring site

Data Understanding

Dataset has 420768 records and 18 features. It contains time-series data. Also it has 53728 N/A values.

Data Preparation

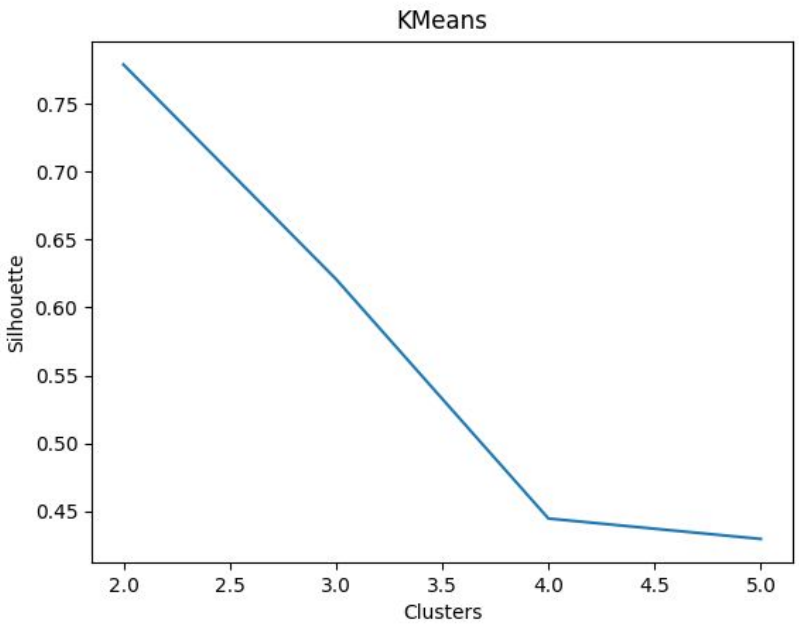
1. Sort data
2. Reduce dimensionality
3. Drop useless features
4. Make slides
5. Transform data

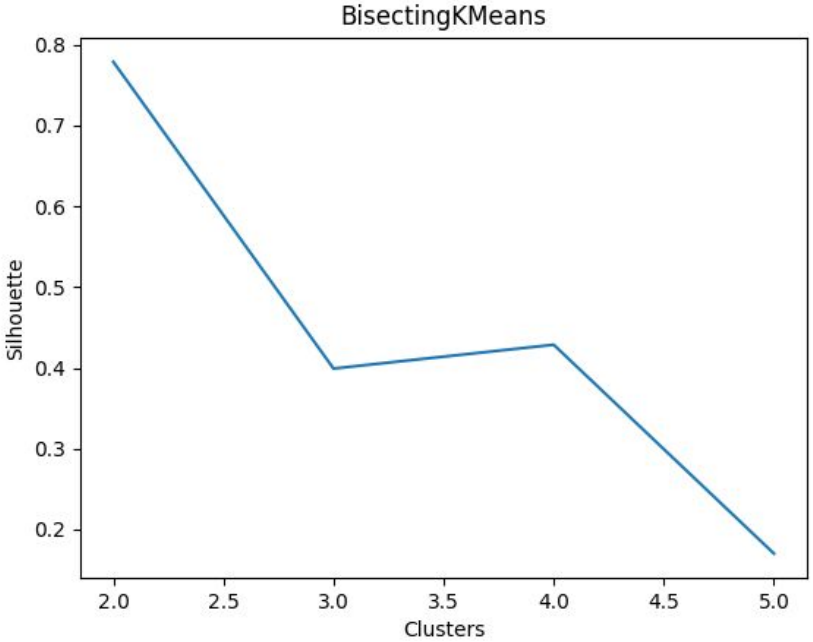
Modeling

I preferred models that work on the basis of centroids, because later, with new data, you can very quickly determine which cluster it belongs to. I need only to save coordinates of centroids.

1. K-Means
2. Bisecting K-Means

Evaluation


KMeans
Silhouette
0.75
0.70
0.65
0.60
0.55
0.50
0.45
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Clusters


BisectingKMeans
Silhouette
0.8
0.7
0.6
0.5
0.4
0.3
0.2
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Clusters

Deployment and future

1. Explore clusters with experts in environmental problems.
2. Determine what conditions will be imposed on enterprises.
3. Analyze new real life examples to scale solution.