Cluster Analysis in Spark

Problem Description

This activity guides you through the process of performing cluster analysis on a dataset using k-means. In this activity, we will perform cluster analysis on the *minute-weather.csv* dataset using the k-means algorithm. Recall that this dataset contains weather measurements such as temperature, relative humidity, etc., from a weather station in San Diego, California, collected at one-minute intervals. The goal of cluster analysis on this data is to identify different weather patterns for this weather station.

Learning Objectives

By the end of this activity, you will be able to:

1. Scale all features so that each feature is zero-normalized
2. Create an "elbow" plot, the number of clusters vs. within-cluster sum-of-squared errors, to determine a value for k, the number of clusters in k-means
3. Perform cluster analysis on a dataset using k-means
4. Create parallel coordinates plots to analyze cluster centers

In this activity, you will be programming in a Jupyter Python Notebook. If you have not already started the Jupyter Notebook server, see the instructions in the Reading *Instructions for Starting Jupyter*.

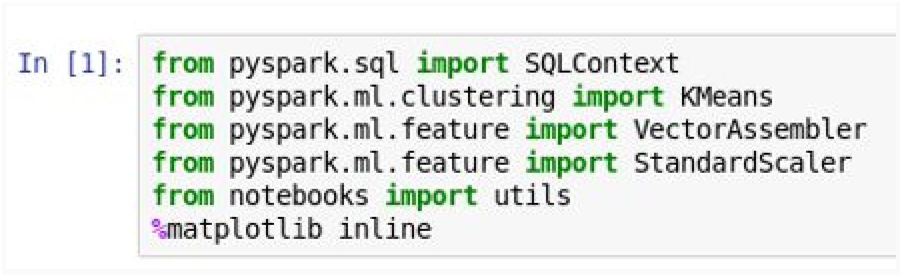
Step 1. **Open Jupyter Python Notebook.** Open a web browser by clicking on the web browser icon at the top of the toolbar:

Navigate to *localhost:8889/tree/Downloads/big-data-4*:

localhost:8889/tree/Downloads/big-data-4

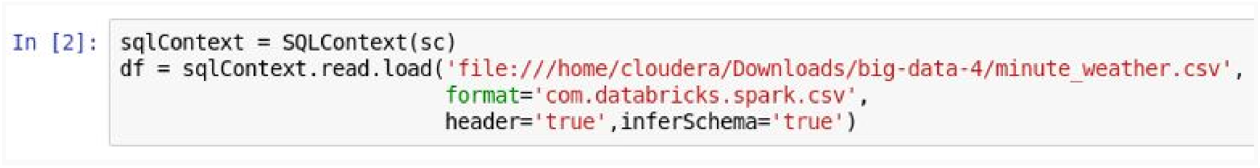
Open the clustering notebook by clicking on *clustering.ipynb:*

Step 2. **Load minute weather data.** Execute the first cell to load the classes used in this activity:



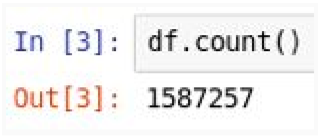
```
In [1]: from pyspark.sql import SQLContext
        from pyspark.ml.clustering import KMeans
        from pyspark.ml.feature import VectorAssembler
        from pyspark.ml.feature import StandardScaler
        from notebooks import utils
        %matplotlib inline
```

Execute the second cell to load the minute weather data in *minute_weather.csv:*



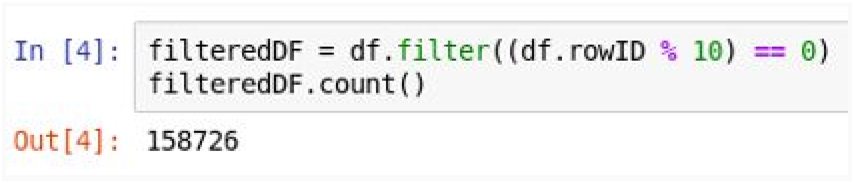
```
In [2]: sqlContext = SQLContext(sc)
        df = sqlContext.read.load('file:///home/cloudera/Downloads/big-data-4/minute_weather.csv',
                                  format='com.databricks.spark.csv',
                                  header='true',inferSchema='true')
```

Step 3. **Subset and remove unused data**. Let's count the number of rows in the DataFrame:



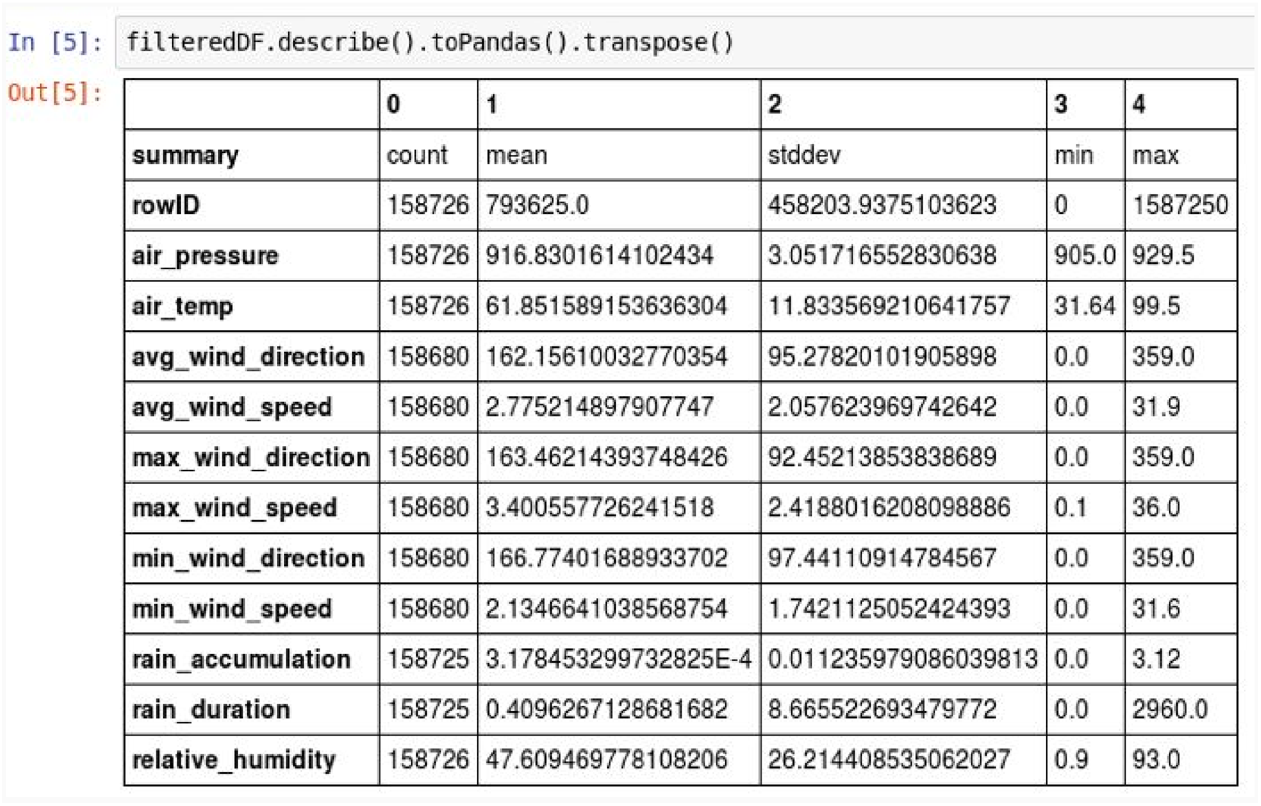
```
In [3]: df.count()

Out[3]: 1587257
```

There are over 1.5 million rows in the DataFrame. Clustering this data on your computer in the Cloudera VM can take a long time, so let's only one-tenth of the data. We can subset by calling *filter()* and using the *rowID* column:



```
In [4]: filteredDF = df.filter((df.rowID % 10) == 0)
        filteredDF.count()

Out[4]: 158726
```

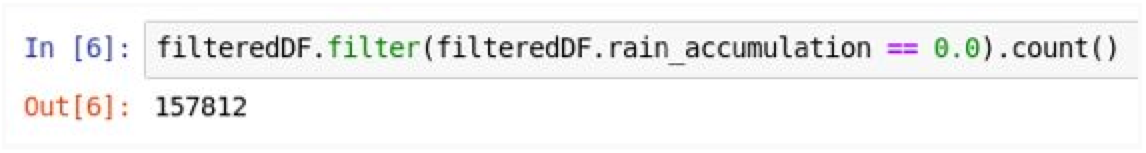
Let's compute the summary statistics using *describe():*



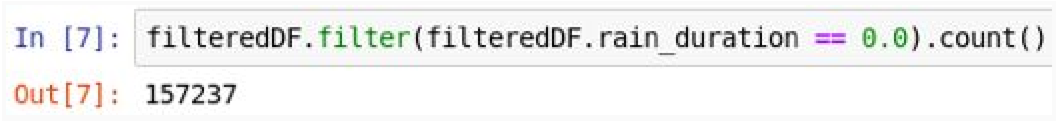
```
In [5]: filteredDF.describe().toPandas().transpose()
Out[5]:
```

	0	1	2	3	4
summary	count	mean	stddev	min	max
rowID	158726	793625.0	458203.9375103623	0	1587250
air_pressure	158726	916.8301614102434	3.051716552830638	905.0	929.5
air_temp	158726	61.851589153636304	11.833569210641757	31.64	99.5
avg_wind_direction	158680	162.15610032770354	95.27820101905898	0.0	359.0
avg_wind_speed	158680	2.775214897907747	2.057623969742642	0.0	31.9
max_wind_direction	158680	163.46214393748426	92.45213853838689	0.0	359.0
max_wind_speed	158680	3.400557726241518	2.4188016208098886	0.1	36.0
min_wind_direction	158680	166.77401688933702	97.44110914784567	0.0	359.0
min_wind_speed	158680	2.1346641038568754	1.7421125052424393	0.0	31.6
rain_accumulation	158725	3.178453299732825E-4	0.011235979086039813	0.0	3.12
rain_duration	158725	0.4096267128681682	8.665522693479772	0.0	2960.0
relative_humidity	158726	47.609469778108206	26.214408535062027	0.9	93.0

The weather measurements in this dataset were collected during a drought in San Diego. We can count the how many values of rain accumulation and duration are 0:

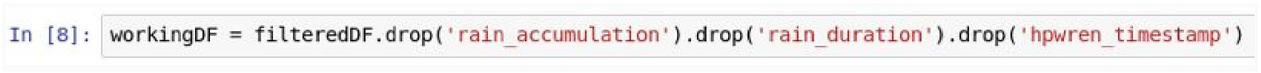


```
In [6]: filteredDF.filter(filteredDF.rain_accumulation == 0.0).count()
Out[6]: 157812
```



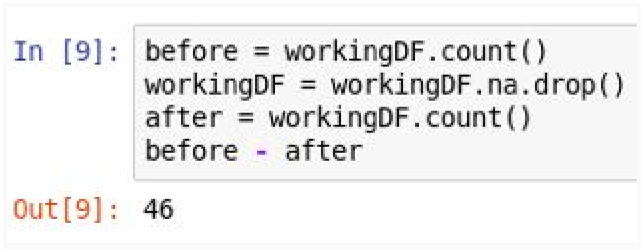
```
In [7]: filteredDF.filter(filteredDF.rain_duration == 0.0).count()
Out[7]: 157237
```

Since most the values for these columns are 0, let's drop them from the DataFrame to speed up our analyses. We can also drop the *hpwren_timestamp* column since we do not use it.



```
In [8]: workingDF = filteredDF.drop('rain_accumulation').drop('rain_duration').drop('hpwren_timestamp')
```

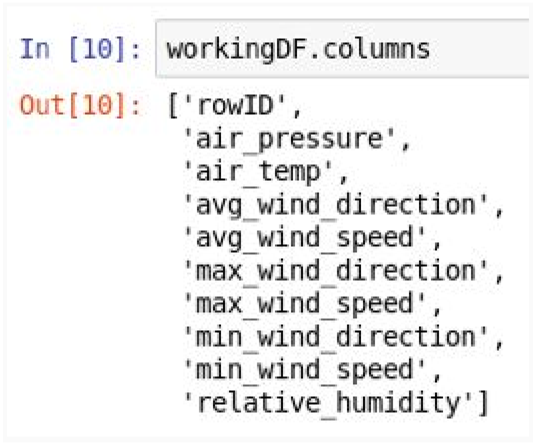
Let's drop rows with missing values and count how many rows were dropped:



```
In [9]:  before = workingDF.count()
          workingDF = workingDF.na.drop()
          after = workingDF.count()
          before - after

Out[9]:   46
```

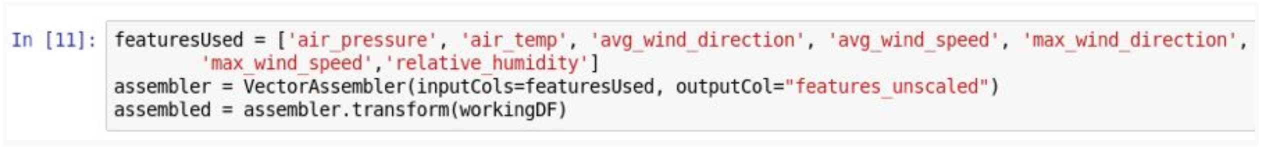
Step 4. **Scale the data**. Since the features are on different scales (e.g., air pressure values are in the 900’s, while relative humidities range from 0 to 100), they need to be scaled. We will scale them so that each feature will have a value of 0 for the mean, and a value of 1 for the standard deviation.

First, we will combine the columns into a single vector column. Let's look at the columns in the DataFrame:



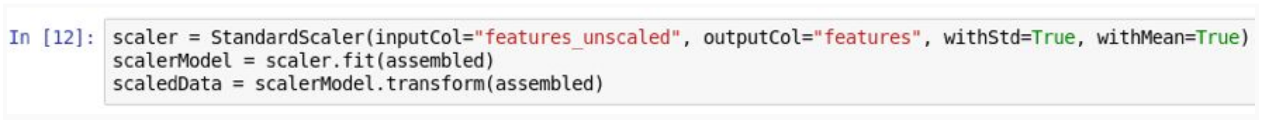
```
In [10]:  workingDF.columns

Out[10]:  ['rowID',
           'air_pressure',
           'air_temp',
           'avg_wind_direction',
           'avg_wind_speed',
           'max_wind_direction',
           'max_wind_speed',
           'min_wind_direction',
           'min_wind_speed',
           'relative_humidity']
```

We do not want to include *rowID* since it is the row number. The minimum wind measurements have a high correlation to the average wind measurements, so we will not include them either. Let's create an array of the columns we want to combine, and use *VectorAssembler* to create the vector column:



```
In [11]:  featuresUsed = ['air_pressure', 'air_temp', 'avg_wind_direction', 'avg_wind_speed', 'max_wind_direction',
                  'max_wind_speed','relative_humidity']
          assembler = VectorAssembler(inputCols=featuresUsed, outputCol="features_unscaled")
          assembled = assembler.transform(workingDF)
```

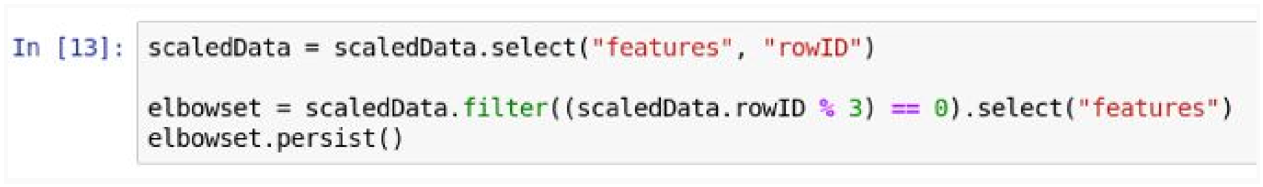
Next, let's use *StandardScaler* to scale the data:



```
In [12]:  scaler = StandardScaler(inputCol="features_unscaled", outputCol="features", withStd=True, withMean=True)
          scalerModel = scaler.fit(assembled)
          scaledData = scalerModel.transform(assembled)
```

The *withMean* argument specifies to center the data with the mean before scaling, and *withStd* specifies to scale the data to the unit standard deviation.

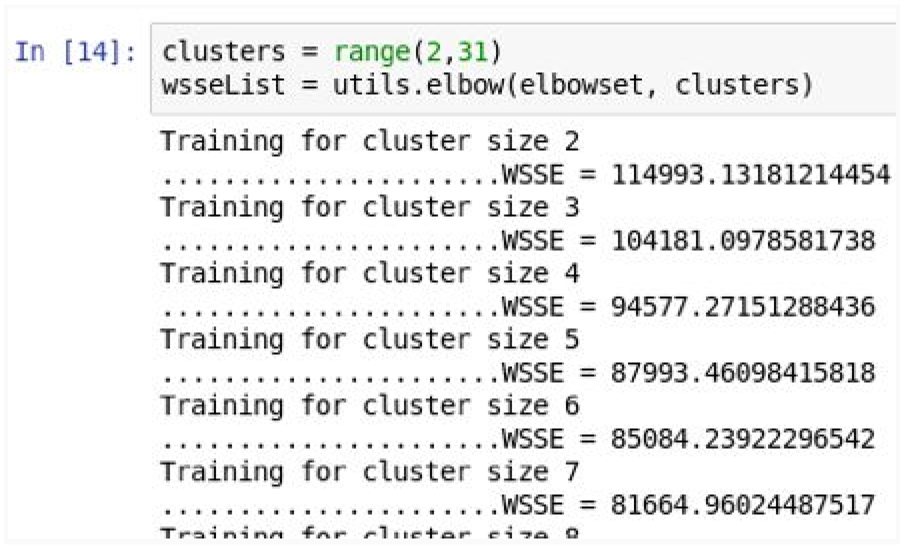
Step 5. **Create elbow plot.** The k-means algorithm requires that the value of k, the number of clusters, to be specified. To determine a good value for *k*, we will use the “elbow” method. This method involves applying k-means, using different values for *k*, and calculating the within-cluster sum-of-squared error (WSSE). Since this means applying k-means multiple times, this process can be very compute-intensive. To speed up the process, we will use only a subset of the dataset. We will take every third sample from the dataset to create this subset:



```
In [13]: scaledData = scaledData.select("features", "rowID")

         elbowset = scaledData.filter((scaledData.rowID % 3) == 0).select("features")
         elbowset.persist()
```

The last line calls the *persist()* method to tell Spark to keep the data in memory (if possible), which will speed up the computations.

Let's compute the k-means clusters for *k* = 2 to 30 to create an elbow plot:

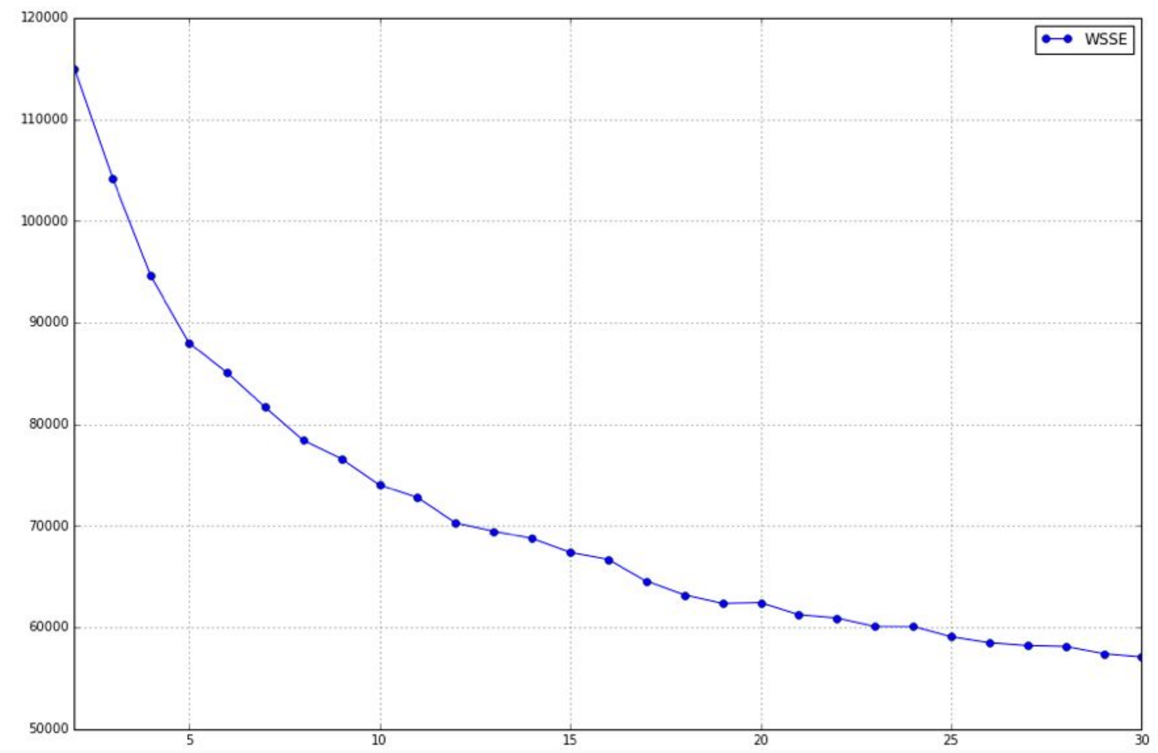


```
In [14]: clusters = range(2,31)
         wsseList = utils.elbow(elbowset, clusters)

         Training for cluster size 2
         ......................WSSE = 114993.13181214454
         Training for cluster size 3
         ......................WSSE = 104181.0978581738
         Training for cluster size 4
         ......................WSSE = 94577.27151288436
         Training for cluster size 5
         ......................WSSE = 87993.46098415818
         Training for cluster size 6
         ......................WSSE = 85084.23922296542
         Training for cluster size 7
         ......................WSSE = 81664.96024487517
         Training for cluster size 8
```

The first line creates an array with the numbers 2 through 30, and the second line calls the *elbow()* function defined in the *utils.py* library to perform clustering. The first argument to *elbow()* is the dataset, and the second is the array of values for *k*. The *elbow()* function returns an array of the WSSE for each number of clusters.

Let's plot the results by calling *elbow_plot()* in the *utils.py* library:

```
In [15]: utils.elbow_plot(wsseList, clusters)
```



The values for *k* are plotted against the WSSE values, and the elbow, or bend in the curve, provides a good estimate for the value for *k*. In this plot, we see that the elbow in the curve is between 10 and 15, so let's choose *k* = 12. We will use this value to set the number of clusters for k-means.

Step 6. **Cluster using selected k**. Let's select the data we want to cluster:

```
In [16]: scaledDataFeat = scaledData.select("features")
         scaledDataFeat.persist()
```

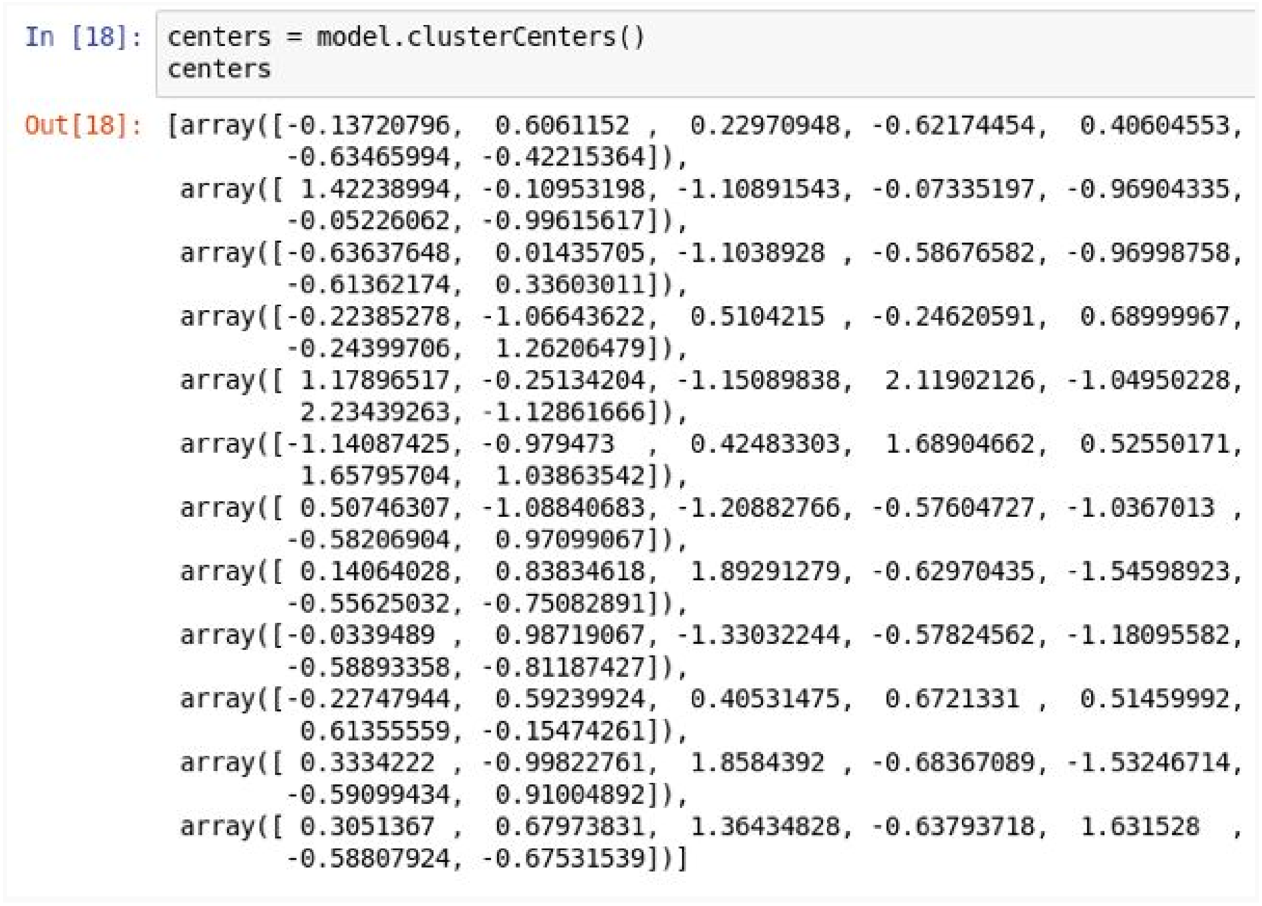
Again, we call the *persist()* method to cache the data in memory for faster access.

We can perform clustering using *KMeans:*

```
In [17]: kmeans = KMeans(k=12, seed=1)
         model = kmeans.fit(scaledDataFeat)
         transformed = model.transform(scaledDataFeat)
```

The first line creates a new *KMeans* instance with 12 clusters and a specific seed value. (As in previous hands-on activities, we use a specific seed value for reproducible results.) The second line fits the data to the model, and the third applies the model to the data.

Once the model is created, we can determine the center measurement of each cluster:

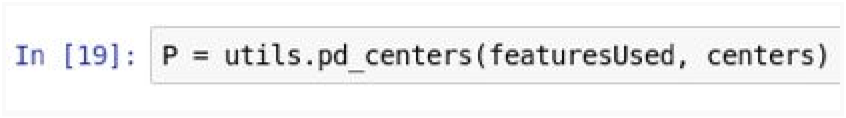


```
In [18]: centers = model.clusterCenters()
         centers

Out[18]: [array([-0.13720796,  0.6061152 ,  0.22970948, -0.62174454,  0.40604553,
                -0.63465994, -0.42215364]),
          array([ 1.42238994, -0.10953198, -1.10891543, -0.07335197, -0.96904335,
                -0.05226062, -0.99615617]),
          array([-0.63637648,  0.01435705, -1.1038928 , -0.58676582, -0.96998758,
                -0.61362174,  0.33603011]),
          array([-0.22385278, -1.06643622,  0.5104215 , -0.24620591,  0.68999967,
                -0.24399706,  1.26206479]),
          array([ 1.17896517, -0.25134204, -1.15089838,  2.11902126, -1.04950228,
                 2.23439263, -1.12861666]),
          array([-1.14087425, -0.979473  ,  0.42483303,  1.68904662,  0.52550171,
                 1.65795704,  1.03863542]),
          array([ 0.50746307, -1.08840683, -1.20882766, -0.57604727, -1.0367013 ,
                -0.58206904,  0.97099067]),
          array([ 0.14064028,  0.83834618,  1.89291279, -0.62970435, -1.54598923,
                -0.55625032, -0.75082891]),
          array([-0.0339489 ,  0.98719067, -1.33032244, -0.57824562, -1.18095582,
                -0.58893358, -0.81187427]),
          array([-0.22747944,  0.59239924,  0.40531475,  0.6721331 ,  0.51459992,
                 0.61355559, -0.15474261]),
          array([ 0.3334222 , -0.99822761,  1.8584392 , -0.68367089, -1.53246714,
                -0.59099434,  0.91004892]),
          array([ 0.3051367 ,  0.67973831,  1.36434828, -0.63793718,  1.631528  ,
                -0.58807924, -0.67531539])]
```

It is difficult to compare the cluster centers by just looking at these numbers. So we will use plots in the next step to visualize them.

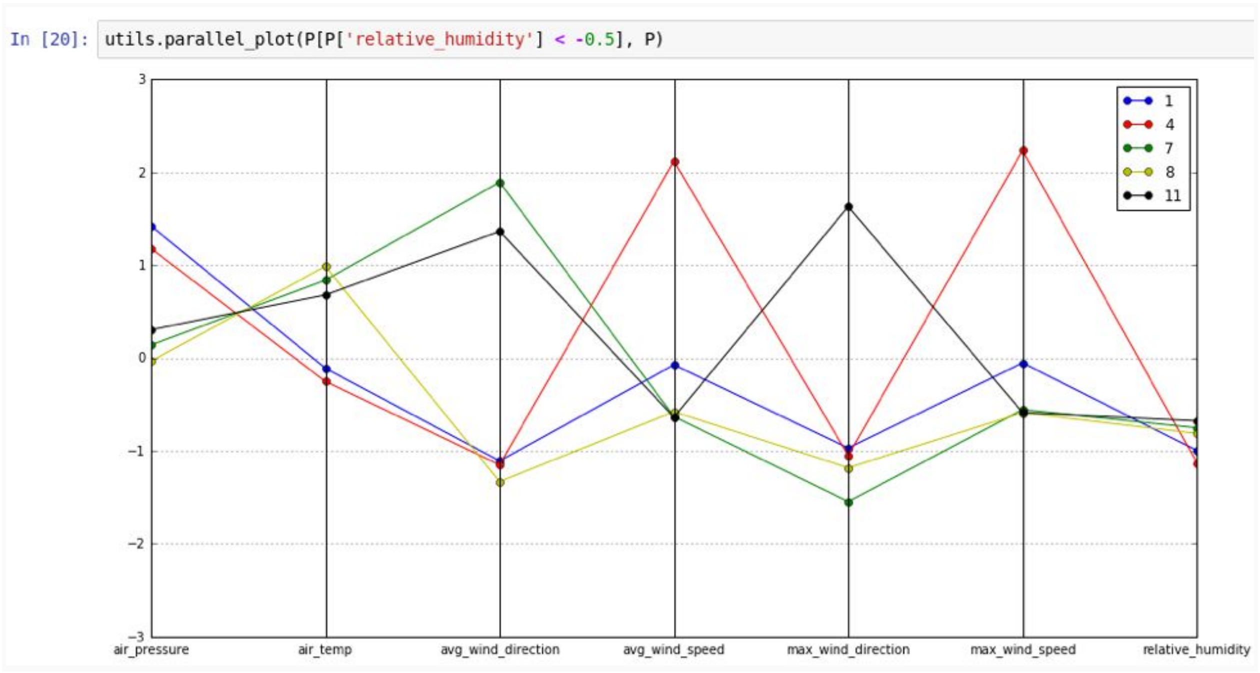
Step 7. **Create parallel plots of clusters and analysis.** A parallel coordinates plot is a great way to visualize multi-dimensional data. Each line plots the centroid of a cluster, and all of the features are plotted together. Recall that the feature values were scaled to have mean = 0 and standard deviation = 1. So the values on the y-axis of these parallel coordinates plots show the number of standard deviations from the mean. For example, +1 means one standard deviation higher than the mean of all samples, and -1 means one standard deviation lower than the mean of all samples.

We'll create the plots with *matplotlib* using a Pandas DataFrame each row contains the cluster center coordinates and cluster label. (Matplotlib can plot Pandas DataFrames, but not Spark DataFrames.) Let's use the *pd_centers()*function in the *utils.py* library to create the Pandas DataFrame:



```
In [19]: P = utils.pd_centers(featuresUsed, centers)
```

Let's show clusters for "Dry Days", i.e., weather samples with low relative humidity:



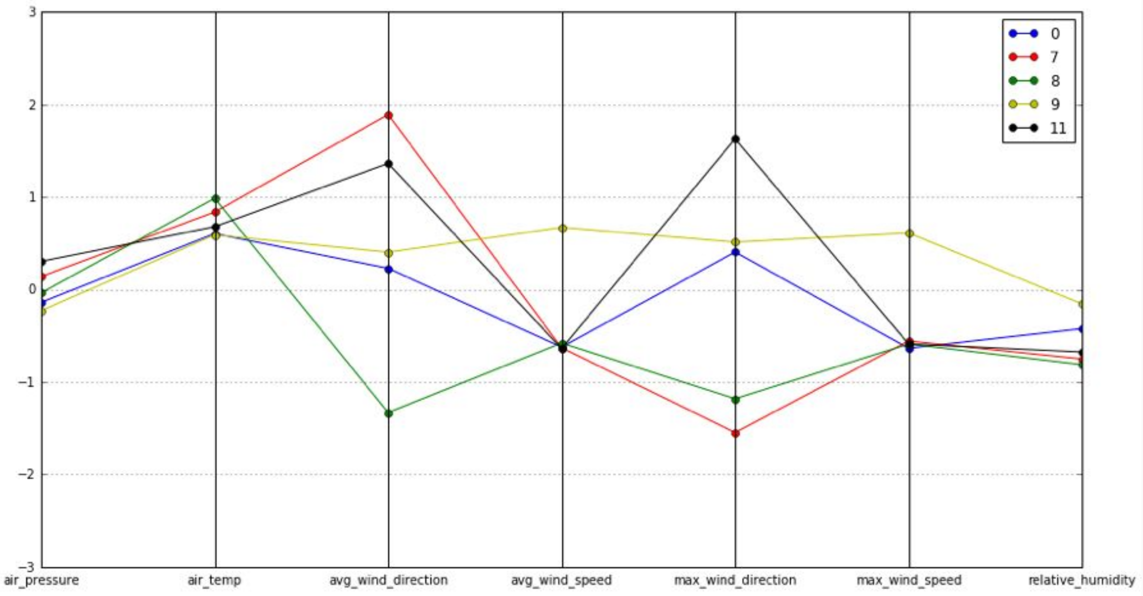
```
In [20]: utils.parallel_plot(P[P['relative_humidity'] < -0.5], P)
```

The first argument to *parallel_plot* selects the clusters whose relative humidities are centered less than 0.5 from the mean value. All clusters in this plot have *relative_humidity* < -0.5, but they differ in values for other features, meaning that there are several weather patterns that include low humidity.

Note in particular cluster 4. This cluster has samples with lower-than-average wind direction values. Recall that wind direction values are in degrees, and 0 means wind coming from the North and increasing clockwise. So samples in this cluster have wind coming from the N and NE directions, with very high wind speeds, and low relative humidity. These are characteristic weather patterns for Santa Ana conditions, which greatly increase the dangers of wildfires.

Let's show clusters for "Warm Days", i.e., weather samples with high air temperature:

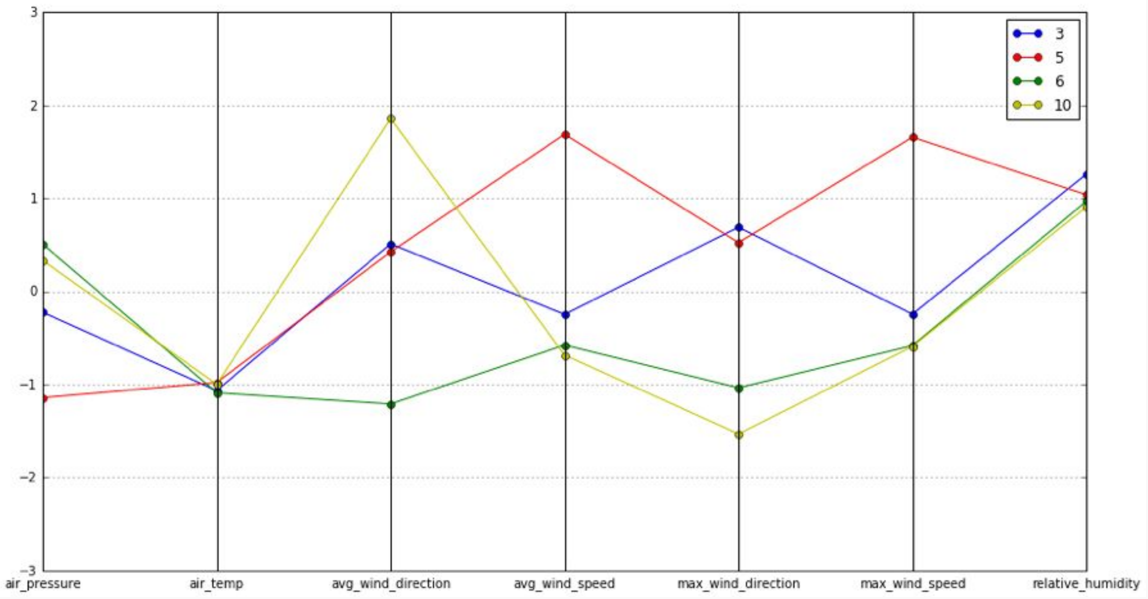
```
In [21]: utils.parallel_plot(P[P['air_temp'] > 0.5], P)
```



All clusters in this plot have *air_temp* > 0.5, but they differ in values for other features.

Let's show clusters for "Cool Days", i.e., weather samples with high relative humidity and low air temperature:

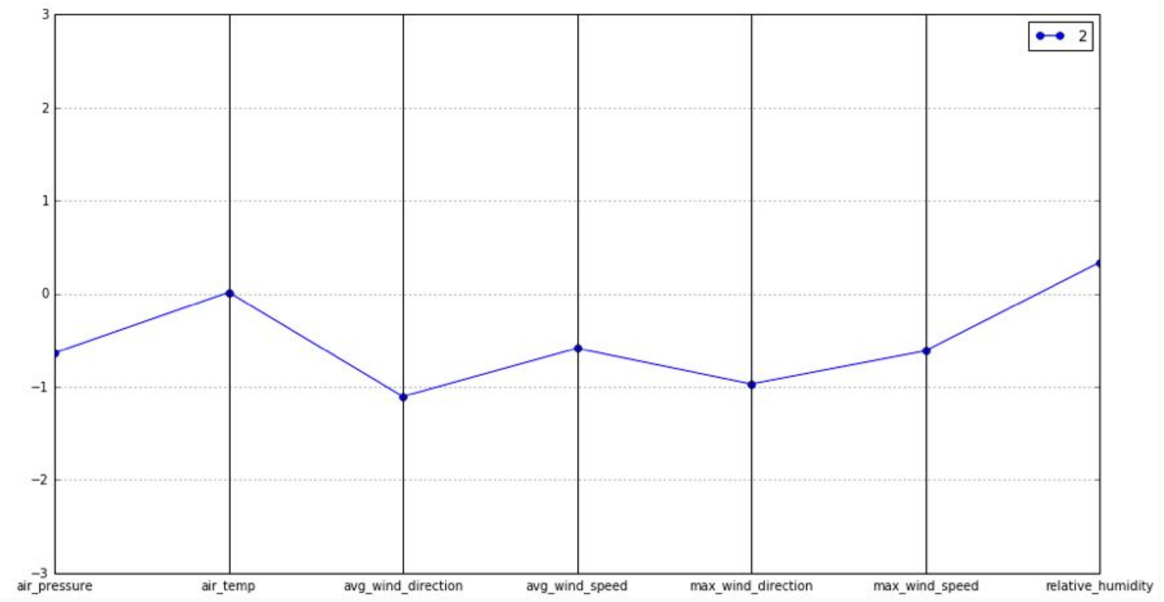
```
In [22]: utils.parallel_plot(P[(P['relative_humidity'] > 0.5) & (P['air_temp'] < 0.5)], P)
```



All clusters in this plot have *relative_humidity* > 0.5 and *air_temp* < 0.5. These clusters represent cool temperature with high humidity and possibly rainy weather patterns. For cluster 5, note that the wind speed values are high, suggesting stormy weather patterns with rain and wind.

So far, we've seen all the clusters except 2 since it did not fall into any of the other categories. Let's plot this cluster:

```
In [23]: utils.parallel_plot(P.iloc[[2]], P)
```



Cluster 2 captures days with mild weather.