9-FINAL

November 26, 2018

1 9 Clustering

1.1 1. DBSCAN

Using DBSCAN iterate (for-loop) through different values of `min_samples` (1 to 10) and `epsilon` (.05 to .5, in steps of .01) to find clusters in the road-data used in the Lesson and calculate the Silohouette Coeff for `min_samples` and `epsilon`. Plot ***one*** line plot with the multiple lines generated from the min_samples and epsilon values. Use a 2D array to store the SilCoeff values, one dimension represents `min_samples`, the other represents epsilon.

```
In [1]: import pandas as pd
        # allow plots to appear in the notebook
        %matplotlib notebook
        import matplotlib.pyplot as plt
        import seaborn
        from mpl_toolkits.mplot3d import Axes3D
        plt.rcParams['font.size'] = 14
        # plt.rcParams['figure.figsize'] = (20.0, 10.0)
```

```
In [8]: X= pd.read_csv('../data/3D_spatial_network.csv')
        X = X.drop(['osm'], axis=1).sample(20000)
        X.head()
```

```
Out[8]:               lat        lon        alt
        101921   8.884733  57.074968  12.761273
        105722   9.501151  56.618347  36.349233
        115070  10.397148  57.583213   4.934445
        162484   9.853012  57.485098  26.427404
        154551  10.054684  56.952023   7.721160
```

```
In [10]: fig = plt.figure()
         X.lat.hist(bins=1000)
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x20a24d31940>

In [11]: XX = X.copy()
         XX['alt'] = (X.alt - X.alt.mean())/X.alt.std()
         XX['lat'] = (X.lat - X.lat.mean())/X.lat.std()
         XX['lon'] = (X.lon - X.lon.mean())/X.lon.std()

In [12]: fig = plt.figure()
         XX.lat.hist(bins=1000)

<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>


Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x20a260fd6a0>

In [13]: fig = plt.figure()
         plt.scatter(XX.lon, XX.lat, alpha=.1, s=5, )

<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>


Out[13]: <matplotlib.collections.PathCollection at 0x20a26b53550>

In [20]: N = 7
         from sklearn.cluster import KMeans
         km = KMeans(n_clusters=N, random_state=1)
         km.fit(X)
         km = KMeans(n_clusters=N, random_state=1)
         XX['cluster'] = km.fit_predict(XX[['lon', 'lat', 'alt']])

In [22]: XX

Out[22]:              lat       lon       alt  cluster
         101921 -1.359060 -0.019020 -0.522747        1
         105722 -0.344782 -1.605500  0.734453        0
         115070  1.129525  1.746826 -0.939904        6
         162484  0.234183  1.405935  0.205636        2
         154551  0.566021 -0.446179 -0.791376        4
         41353  -1.820722 -1.193662 -0.628714        3
         70401   0.393099 -0.131222 -0.825785        4
         147399 -1.554915 -1.242412 -0.011864        3
         46394   0.796073 -1.616820  0.981901        0
         107206 -1.092747 -1.049399 -0.201780        1
         104779 -2.127081 -0.374748  0.154149        3
```

```
103774  1.018504  1.418630  0.413868        2
114326  1.143092  1.876662 -0.528061        6
150030  0.260218 -1.581811  1.750860        0
123040 -0.699981 -0.361896 -0.999061        1
58724  -0.164014  0.495338 -0.756553        4
21230  -0.343537  0.170557 -1.085949        4
84802   0.462884 -0.056620 -1.035927        4
33837  -0.259665 -0.756526 -0.270151        1
136357 -1.655532 -0.707776  2.485934        0
110164  0.799443 -0.514747 -1.001403        4
161369 -1.248852  0.161772 -0.572262        1
148397  0.650556 -0.583682 -0.652734        4
4319   -0.365974 -0.119602 -1.152387        4
64543   0.069241 -1.379018  1.590552        0
159752  1.265767  1.364372 -1.006290        6
65242  -1.434495 -1.002157 -1.175709        1
139996 -0.698047  0.129588  0.570176        1
158495  0.687013  1.255927  0.285765        2
45907   0.626803 -0.269017 -0.568655        4
...          ...       ...       ...      ...
167198  0.471186 -0.158855  0.195361        4
141928 -0.705826 -0.720260 -0.778062        1
95576   0.157445 -1.576606  1.428824        0
55998  -1.114714  0.226163  0.714990        3
154264  1.148113  1.781798 -0.963853        6
5412   -0.690182 -1.070348 -0.270109        1
164457 -2.151345 -1.420887 -1.141300        3
111902 -0.631557 -0.682684 -0.312820        1
27268   0.642718  1.655326 -0.212847        2
37820   1.358720  1.233211 -1.094122        6
155921 -0.218564  0.573231 -0.805502        4
23076   0.279324 -0.223496 -0.118731        4
45195  -1.908462 -1.133282 -1.131066        3
18184   0.604846 -0.905772 -0.055432        4
56936   0.288670 -0.302553 -0.775322        4
95497   0.405097 -0.207907  1.317003        0
160266 -0.081726 -0.297822 -0.864325        4
35000   1.065221  1.073271  2.176458        5
176382 -0.182699  0.617730 -0.219554        4
85032  -2.005120 -0.318751 -0.138097        3
45630   1.267639  1.414138 -0.932710        6
138194 -0.220438 -1.342620  0.958801        0
4709    0.316139 -0.821190  1.905815        0
1754    0.161444 -0.399296  0.057921        4
37323   0.431821  1.722265  0.131768        2
133864 -0.894755 -0.023498 -1.071070        1
92813   0.162857  1.364256  0.142872        2
15028  -0.276577 -1.684381  0.454958        0
```

```
         62089   -0.810209 -0.827578 -0.487991        1
         109157   1.076458  1.965388 -0.443081        6

         [20000 rows x 4 columns]

In [36]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.12)
         XX.cluster = dbscan.fit_predict(XX[['lat','lon']])
         XX.cluster.values_counts()


        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-36-bfe35465c4cc> in <module>()
          2 dbscan = DBSCAN(eps=.12)
          3 XX.cluster = dbscan.fit_predict(XX[['lat','lon']])
    ----> 4 XX.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'


In [38]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(XX['lon'], XX['lat'], c=XX.cluster, s=5, cmap='Paired')

         ax.set_xlabel('lon')
         ax.set_ylabel('lat')
         ax.set_zlabel('alt')
         plt.show()

<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>


In [39]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.12)
         XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
         XX.cluster.values_counts()


        ---------------------------------------------------------------------------
        AttributeError                            Traceback (most recent call last)

        <ipython-input-39-3b3b9f9e8e75> in <module>()
          2 dbscan = DBSCAN(eps=.12)
          3 XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
    ----> 4 XX.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [40]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(XX['lon'], XX['lat'], XX['alt'], c=XX.cluster, s=5, cmap='Paired')

         ax.set_xlabel('lon')
         ax.set_ylabel('lat')
         ax.set_zlabel('alt')
         plt.show()
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>
```

```
In [45]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.1)
         XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
         XX.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-45-43fef2fe1eec> in <module>()
          2 dbscan = DBSCAN(eps=.1)
          3 XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
    ----> 4 XX.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [46]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(XX['lon'], XX['lat'], XX['alt'], c=XX.cluster, s=5, cmap='Paired')

         ax.set_xlabel('lon')
         ax.set_ylabel('lat')
         ax.set_zlabel('alt')
         plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
In [49]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.15)
         XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
         XX.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-49-c2949f6bb1af> in <module>()
          2 dbscan = DBSCAN(eps=.15)
          3 XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
    ----> 4 XX.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [50]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(XX['lon'], XX['lat'], XX['alt'], c=XX.cluster, s=5, cmap='Paired')

         ax.set_xlabel('lon')
         ax.set_ylabel('lat')
         ax.set_zlabel('alt')
         plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
In [54]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.05)
         XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
         XX.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-54-e588c1a374c3> in <module>()
          2 dbscan = DBSCAN(eps=.05)
          3 XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
    ----> 4 XX.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [52]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(XX['lon'], XX['lat'], XX['alt'], c=XX.cluster, s=5, cmap='Paired')

         ax.set_xlabel('lon')
         ax.set_ylabel('lat')
         ax.set_zlabel('alt')
         plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
In [57]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.075)
         XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
         XX.cluster.values_counts()
```

```
    ---------------------------------------------------------------------------

    AttributeError                            Traceback (most recent call last)

    <ipython-input-57-efb622583422> in <module>()
      2 dbscan = DBSCAN(eps=.075)
      3 XX.cluster = dbscan.fit_predict(XX[['lat','lon', 'alt']])
----> 4 XX.cluster.values_counts()


    AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [58]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(XX['lon'], XX['lat'], XX['alt'], c=XX.cluster, s=5, cmap='Paired')

         ax.set_xlabel('lon')
         ax.set_ylabel('lat')
         ax.set_zlabel('alt')
         plt.show()
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>
```

1.2 2. Clustering your own data

Using your own data, find relevant clusters/groups within your data. If your data is labeled already, with a class that you are attempting to predict, be sure to not use it in fitting/training/predicting.

You may use the labels to compare with predictions to show how well the clustering performed using one of the clustering metrics (http://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation).

If you don't have labels, use the silhouette coefficient to show performance. Find the optimal fit for your data but you don't need to be as exhaustive as above.

Additionally, show the clusters in 2D and 3D plots.

For bonus, try using PCA first to condense your data from N columns to less than N.

Two items are expected: - Metric Evaluation Plot - Plots of the clustered data

```
In [85]: beer= pd.read_csv('../data/beers.csv')
         beer= beer.drop(['nid'], axis=1).sample(500)
         beer.head()
```

```
Out[85]:         abv    id                    name                  style  \
        1650  0.065   583          Long Hammer IPA           American IPA
        2126  0.075   122  Golden Frau Honey Wheat                Braggot
        1354  0.056  1907        Montauk Summer Ale     American Blonde Ale
        554   0.050  1219  All American Blonde Ale     American Blonde Ale
        1507  0.053  2112                  Atalanta  Saison / Farmhouse Ale

              brewery_id  ounces
        1650         487    12.0
        2126         282    12.0
        1354         276    12.0
        554          452    12.0
        1507         216    12.0
```

```
In [86]: fig = plt.figure()
         beer.abv.hist(bins=50)
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>


Out[86]: <matplotlib.axes._subplots.AxesSubplot at 0x20a411c0390>
```

```
In [87]: Xbeer = beer.copy()
         Xbeer['brewery_id'] = (beer.brewery_id - beer.brewery_id.mean())/beer.brewery_id.std(
         Xbeer['id'] = (beer.id - beer.id.mean())/beer.id.std()
         Xbeer['abv'] = (beer.abv - beer.abv.mean())/beer.abv.std()
```

```
In [84]: fig = plt.figure()
         plt.scatter(Xbeer.id, Xbeer.abv, alpha=.1, s=5, )
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>


Out[84]: <matplotlib.collections.PathCollection at 0x20a411926a0>
```

```
In [88]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.12)
         Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv']])
         Xbeer.cluster.values_counts()
```

```
C:\Users\Erin\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: UserWarning: Pandas doesn't
  This is separate from the ipykernel package so we can avoid doing imports until
```

```
    ---------------------------------------------------------------------------

    AttributeError                            Traceback (most recent call last)

    <ipython-input-88-030fd458d29e> in <module>()
      2 dbscan = DBSCAN(eps=.12)
      3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv']])
----> 4 Xbeer.cluster.values_counts()


    AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [89]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(Xbeer['id'], Xbeer['abv'], c=Xbeer.cluster, s=5, cmap='Paired')

         ax.set_xlabel('id')
         ax.set_ylabel('abv')
         ax.set_zlabel('ounces')
         plt.show()
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>
```

```
In [92]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.12)
         Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
         Xbeer.cluster.values_counts()
```

```
    ---------------------------------------------------------------------------

    AttributeError                            Traceback (most recent call last)

    <ipython-input-92-4d299e375c8a> in <module>()
      2 dbscan = DBSCAN(eps=.12)
      3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
----> 4 Xbeer.cluster.values_counts()
```

```
        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [93]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(Xbeer['id'], Xbeer['abv'], c=Xbeer.cluster, s=5, cmap='Paired')

         ax.set_xlabel('id')
         ax.set_ylabel('abv')
         ax.set_zlabel('ounces')
         plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
In [94]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.1)
         Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
         Xbeer.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-94-2aea53f2b831> in <module>()
          2 dbscan = DBSCAN(eps=.1)
          3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
    ----> 4 Xbeer.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [95]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(Xbeer['id'], Xbeer['abv'], c=Xbeer.cluster, s=5, cmap='Paired')
         ax.set_xlabel('id')
         ax.set_ylabel('abv')
         ax.set_zlabel('ounces')
         plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
In [96]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.075)
         Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
         Xbeer.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-96-66eb03fa8131> in <module>()
          2 dbscan = DBSCAN(eps=.075)
          3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
    ----> 4 Xbeer.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [97]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(Xbeer['id'], Xbeer['abv'], c=Xbeer.cluster, s=5, cmap='Paired')

         ax.set_xlabel('id')
         ax.set_ylabel('abv')
         ax.set_zlabel('ounces')
         plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
In [98]: from sklearn.cluster import DBSCAN
         dbscan = DBSCAN(eps=.15)
         Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
         Xbeer.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-98-ae87ccad3a4f> in <module>()
          2 dbscan = DBSCAN(eps=.15)
          3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['id', 'abv', 'ounces']])
    ----> 4 Xbeer.cluster.values_counts()


        AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [99]: fig = plt.figure(1)
         plt.clf()
         ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

         plt.cla()

         ax.scatter(Xbeer['id'], Xbeer['abv'], c=Xbeer.cluster, s=5, cmap='Paired')

         ax.set_xlabel('id')
         ax.set_ylabel('abv')
         ax.set_zlabel('ounces')
         plt.show()
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>
```

```
In [100]: from sklearn.cluster import DBSCAN
          dbscan = DBSCAN(eps=.12)
          Xbeer.cluster = dbscan.fit_predict(Xbeer[['brewery_id', 'abv', 'ounces']])
          Xbeer.cluster.values_counts()
```

```
        ---------------------------------------------------------------------------

        AttributeError                            Traceback (most recent call last)

        <ipython-input-100-f44c071142f2> in <module>()
```

```
      2 dbscan = DBSCAN(eps=.12)
      3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['brewery_id', 'abv', 'ounces']])
----> 4 Xbeer.cluster.values_counts()


    AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [101]: fig = plt.figure(1)
          plt.clf()
          ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)

          plt.cla()

          ax.scatter(Xbeer['brewery_id'], Xbeer['abv'], c=Xbeer.cluster, s=5, cmap='Paired')

          ax.set_xlabel('brewery_id')
          ax.set_ylabel('abv')
          ax.set_zlabel('ounces')
          plt.show()
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.HTML object>
```

```
In [104]: from sklearn.cluster import DBSCAN
          dbscan = DBSCAN(eps=.1)
          Xbeer.cluster = dbscan.fit_predict(Xbeer[['brewery_id', 'ounces', 'abv']])
          Xbeer.cluster.values_counts()
```

```
    ---------------------------------------------------------------------------

    AttributeError                            Traceback (most recent call last)

    <ipython-input-104-314acf5ac704> in <module>()
      2 dbscan = DBSCAN(eps=.1)
      3 Xbeer.cluster = dbscan.fit_predict(Xbeer[['brewery_id', 'ounces', 'abv']])
----> 4 Xbeer.cluster.values_counts()


    AttributeError: 'numpy.ndarray' object has no attribute 'values_counts'
```

```
In [105]: fig = plt.figure(1)
          plt.clf()
          ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=140)
```

```
plt.cla()

ax.scatter(Xbeer['brewery_id'], Xbeer['ounces'], c=Xbeer.cluster, s=5, cmap='Paired'

ax.set_xlabel('brewery_id')
ax.set_ylabel('ounces')
ax.set_zlabel('abv')
plt.show()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

1.3 Note

You may use any for both parts 1 and 2, I only recommend using the data I used in the Lesson for part 1. I've included several new datasets in the `data/` folder, such as `beers.csv`, `snow_tweets.csv`, `data/USCensus1990.data.txt.gz`. You do not need to unzip or ungzip any data files. Pandas can open these files on its own.