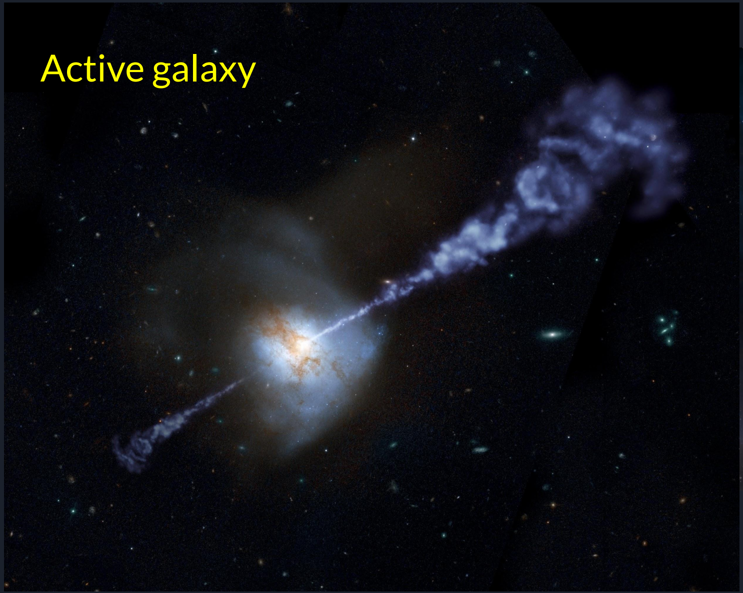
# Outlier Quasars

Reza Monadi
Arman Irani
Hasin Us Sami
Terrance Kuo

# What is a quasar?



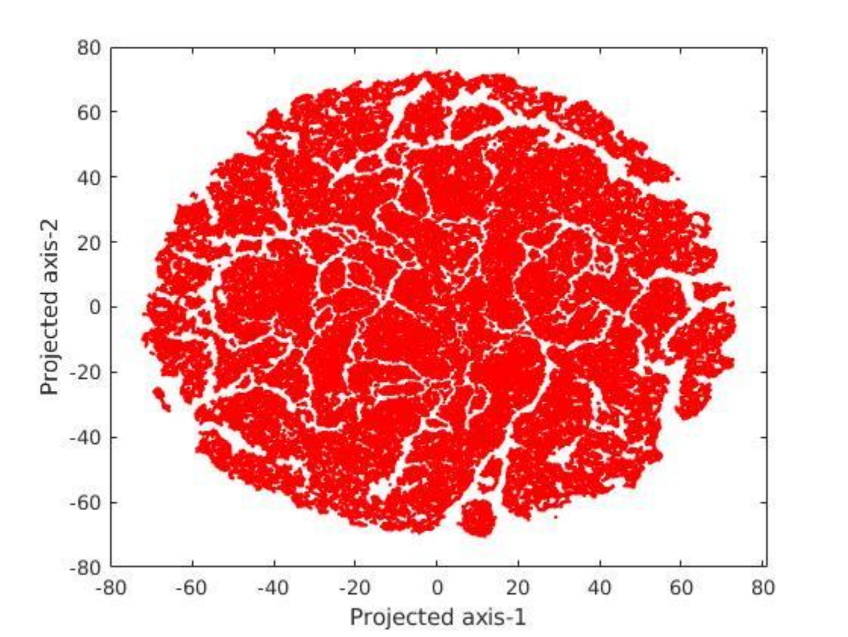
# Data and Preprocessing

- Catalog of quasars (~750,000X180)
- Focus → filtered fluxes
- Flux difference → color of quasars
- Selecting most reliable colors (7 colors)
- Keeping quasars with 2.7<redshift<3
- Removing objects with low Signal/Noise (~#180,000)
- Standard scaling using variance and mean

# Dimensionality Reduction with tSNE

Why tSNE?

- Less susceptible to outliers
- Local, not global like  PCA
- Preserving higher-D structure better in the mapped lower-D



# DBSCAN:

**Key Concepts:**

-A density based clustering algorithm

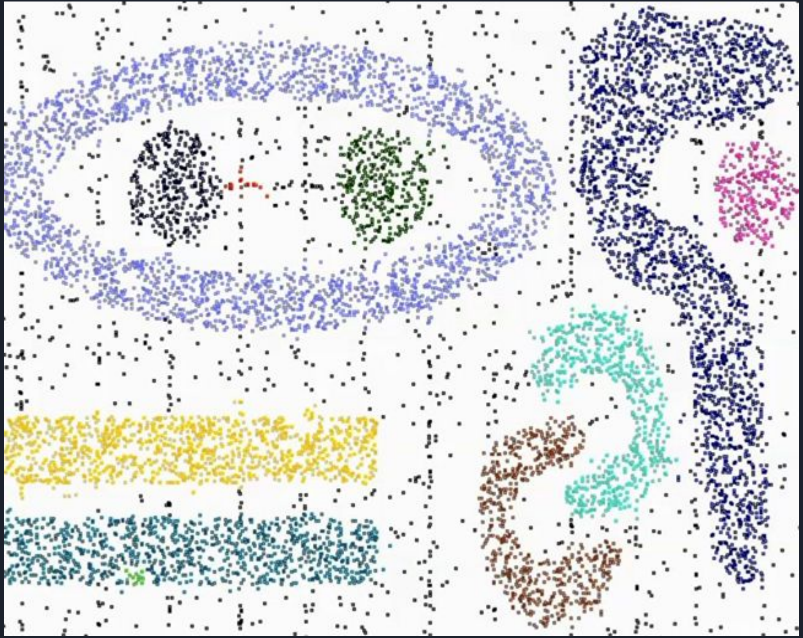
-Insensitive to outliers

-Forms clusters of arbitrary shape

-Does not need the number of clusters to be specified

-Two parameters(epsilon,minimum points) needed

to be specified



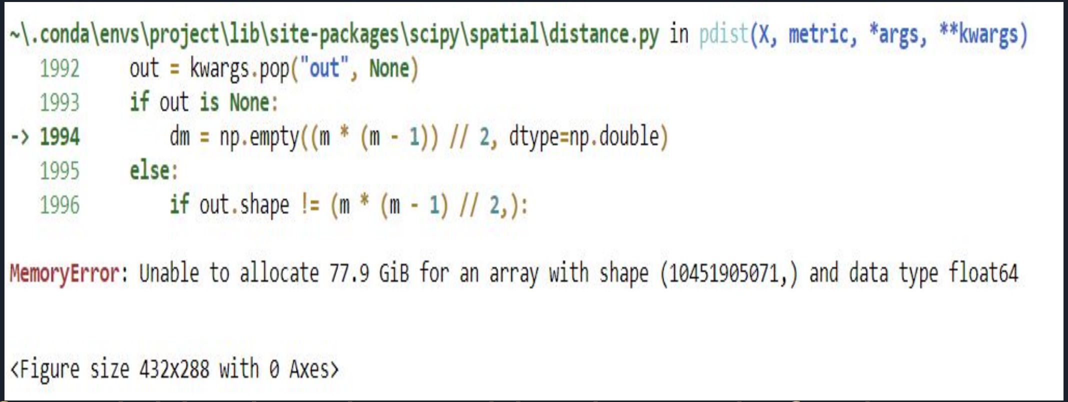
# DBSCAN Implementation from Scratch:

## Challenges:

1)Memory Constraint

2)Run Time

3)selection of parameters



```
~\.conda\envs\project\lib\site-packages\scipy\spatial\distance.py in pdist(X, metric, *args, **kwargs)
   1992     out = kwargs.pop("out", None)
   1993     if out is None:
-> 1994         dm = np.empty((m * (m - 1)) // 2, dtype=np.double)
   1995     else:
   1996         if out.shape != (m * (m - 1) // 2,):

MemoryError: Unable to allocate 77.9 GiB for an array with shape (10451905071,) and data type float64


<Figure size 432x288 with 0 Axes>
```

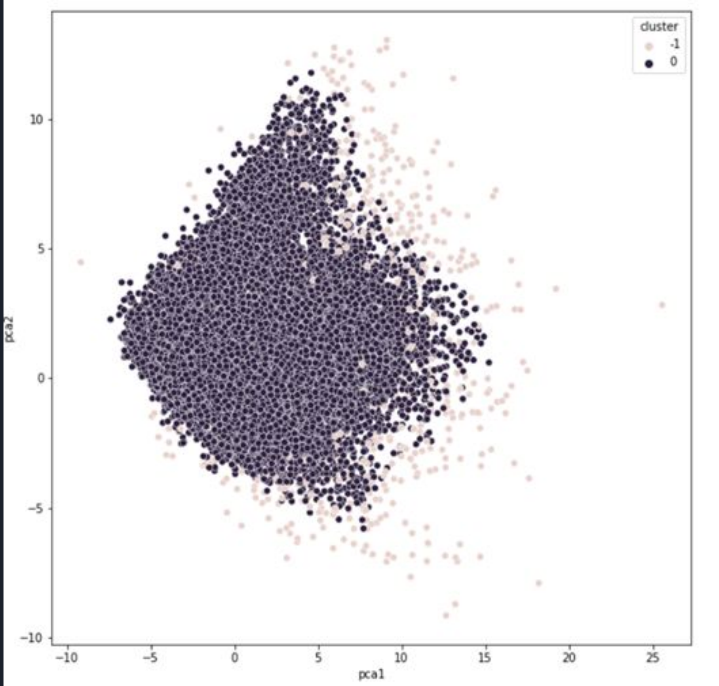
## Solution:

-Using KD-tree approach for neighborhood searching instead of using distance matrix

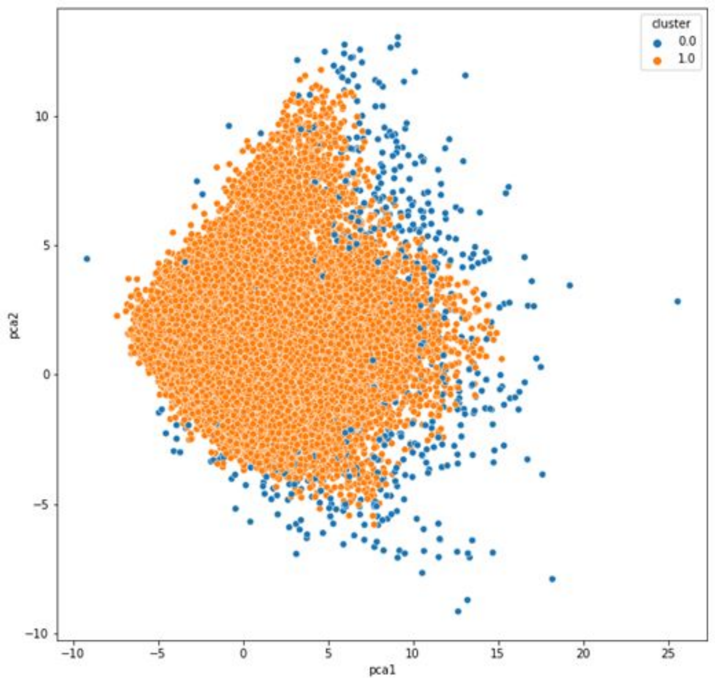
-Using the optimum value for parameters that gives the best isolation of normal points and outliers

# Data Visualization in Reduced Dimension:

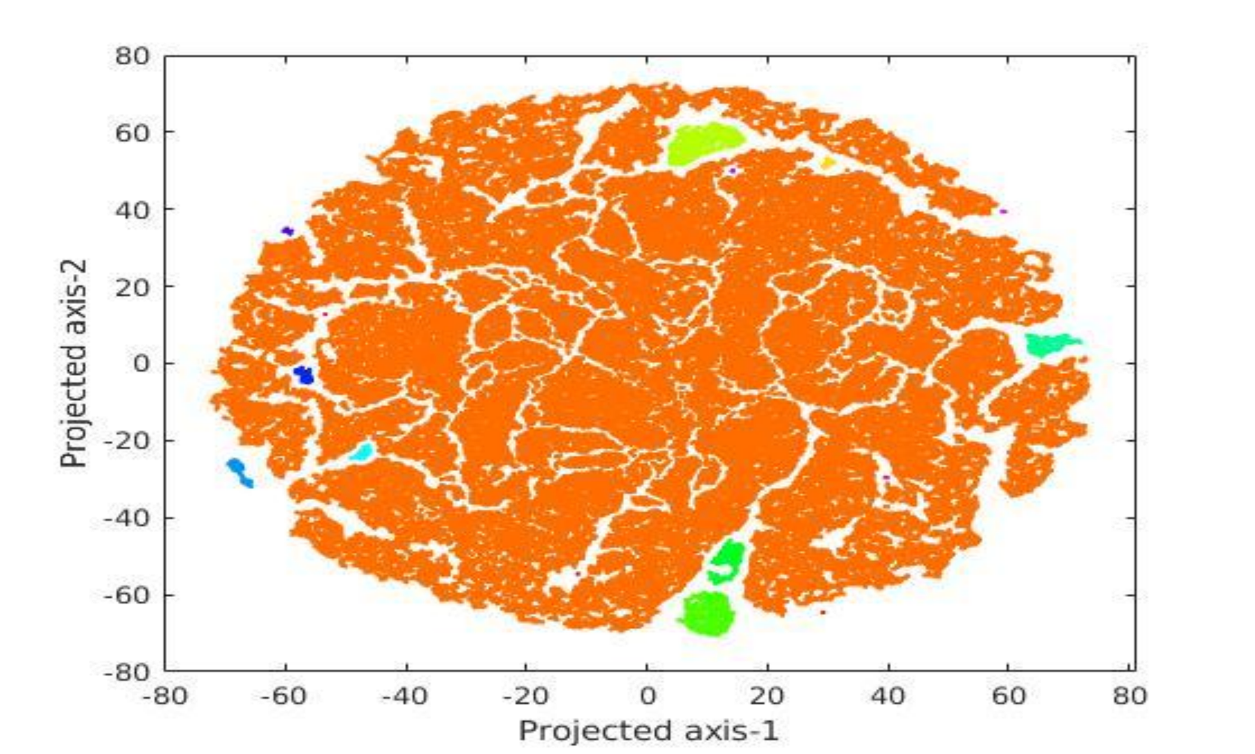
DBSCAN Built-in function:



DBSCAN Implementation from scratch:



# DBSCAN on mapped 2D space

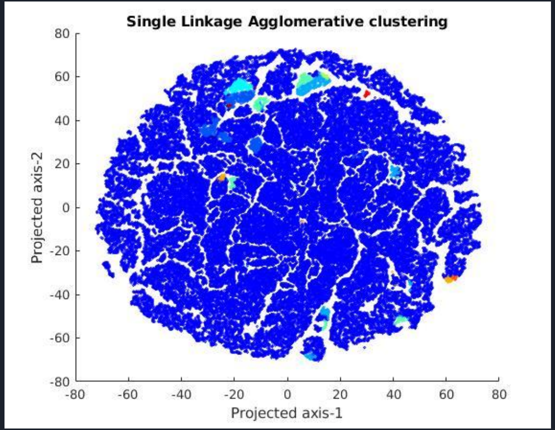


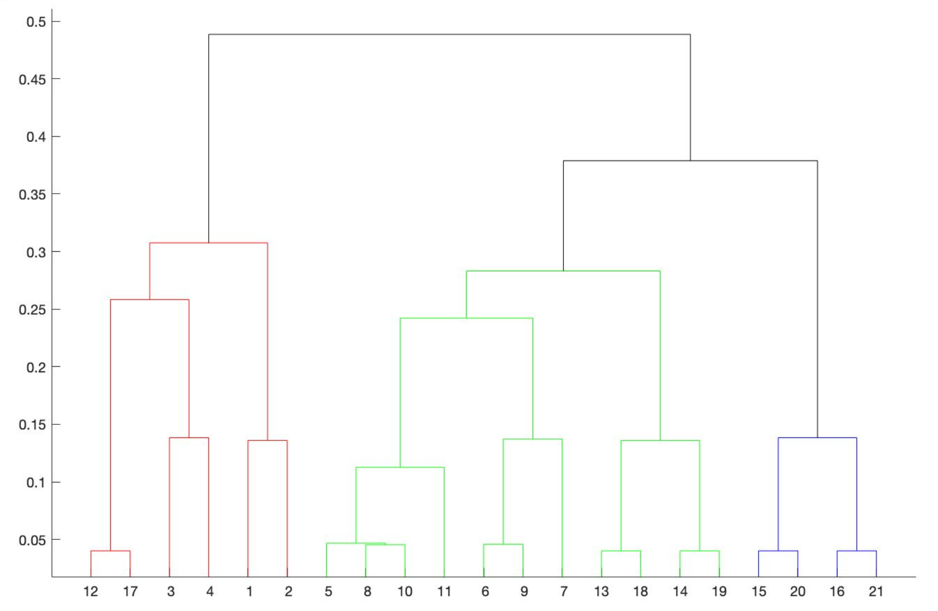
# Agglomerative Clustering

- This lets us cluster our data into groups that has easily visualized results

  Chart on right is our data with reduced dimensions split into 3 clusters.

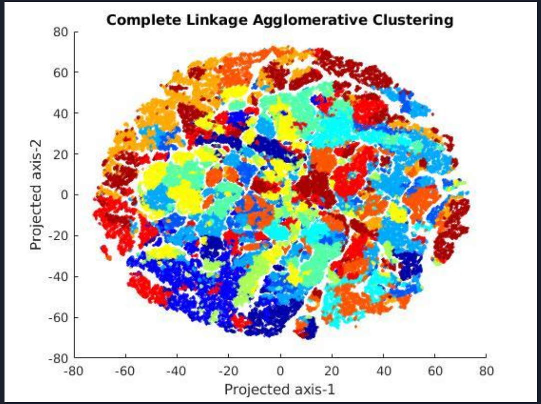
- We can notice that the blue is in a smaller cluster compared to red and green.

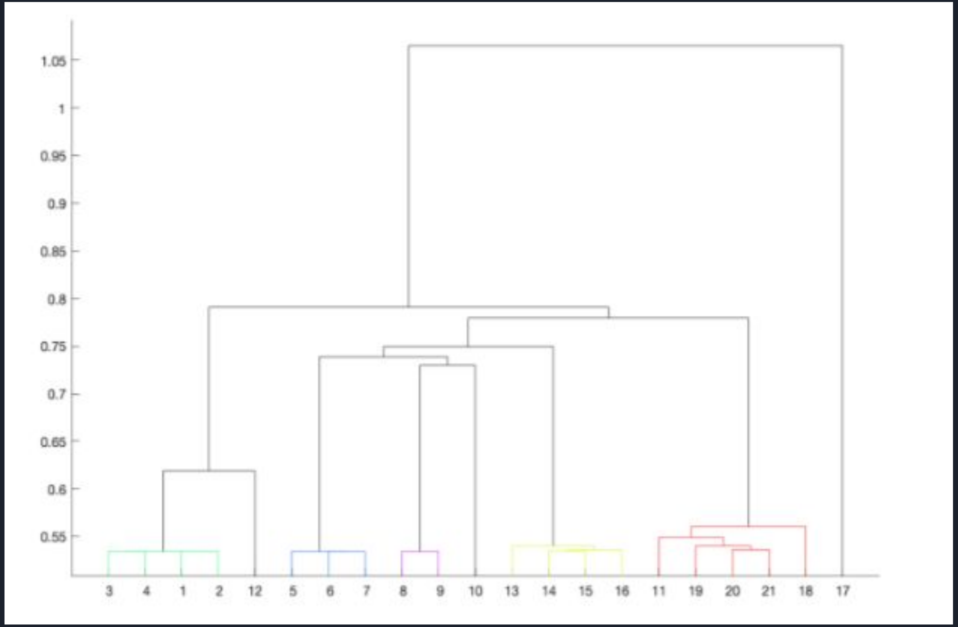




# Agglomerative Clustering

- Another test using Euclidean distance and making the number of clusters 7, we can also visualize the outliers in our data.
- Points 17, 12, and 10 would be outliers in our data due to only having 1 unit in their clusters.



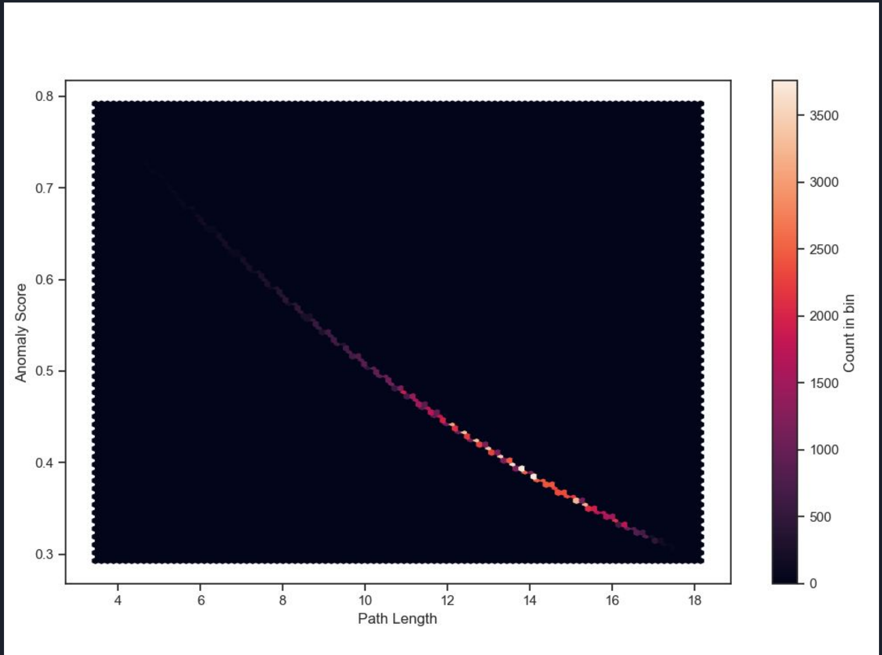


# Isolation Forest

- Outlier detection based on an instance’s ability to be partitioned.

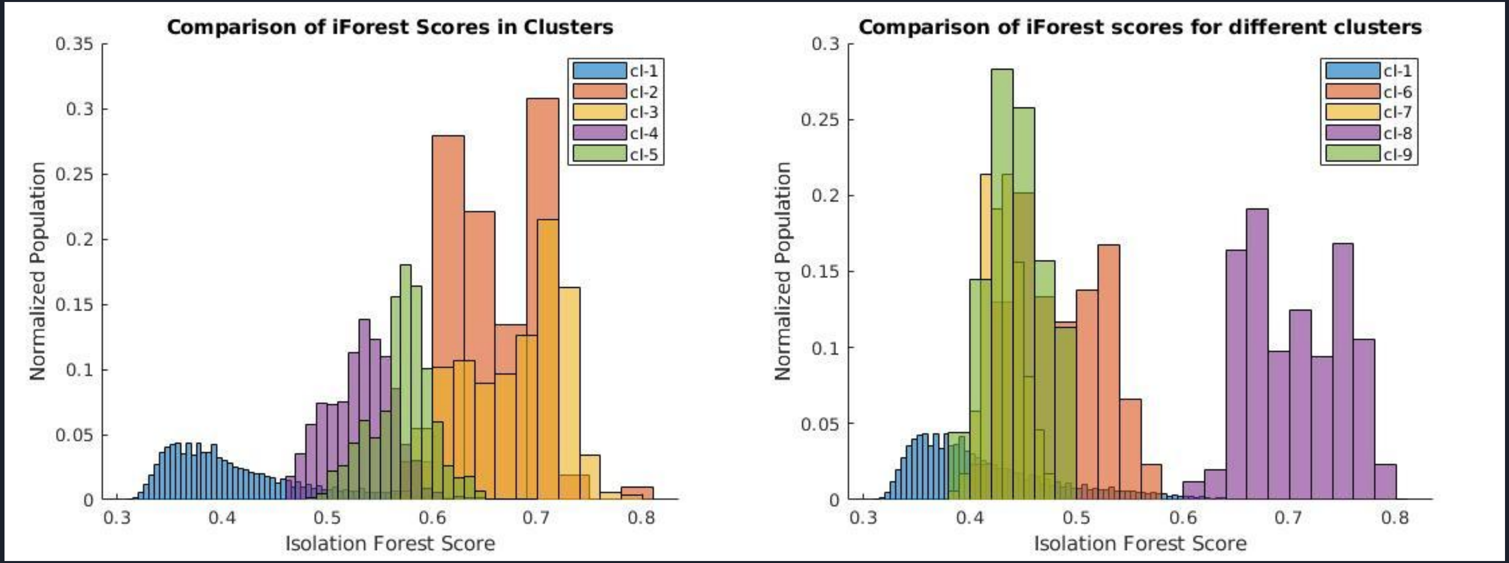
- Assumptions
  - Outlier’s attributes have a high degree of variability from normal.
  - Outlier’s do not occur often.

- Ensemble of Decision Trees
  - *Expected Path Length < Average Path Length = Outlier*



$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}},$$

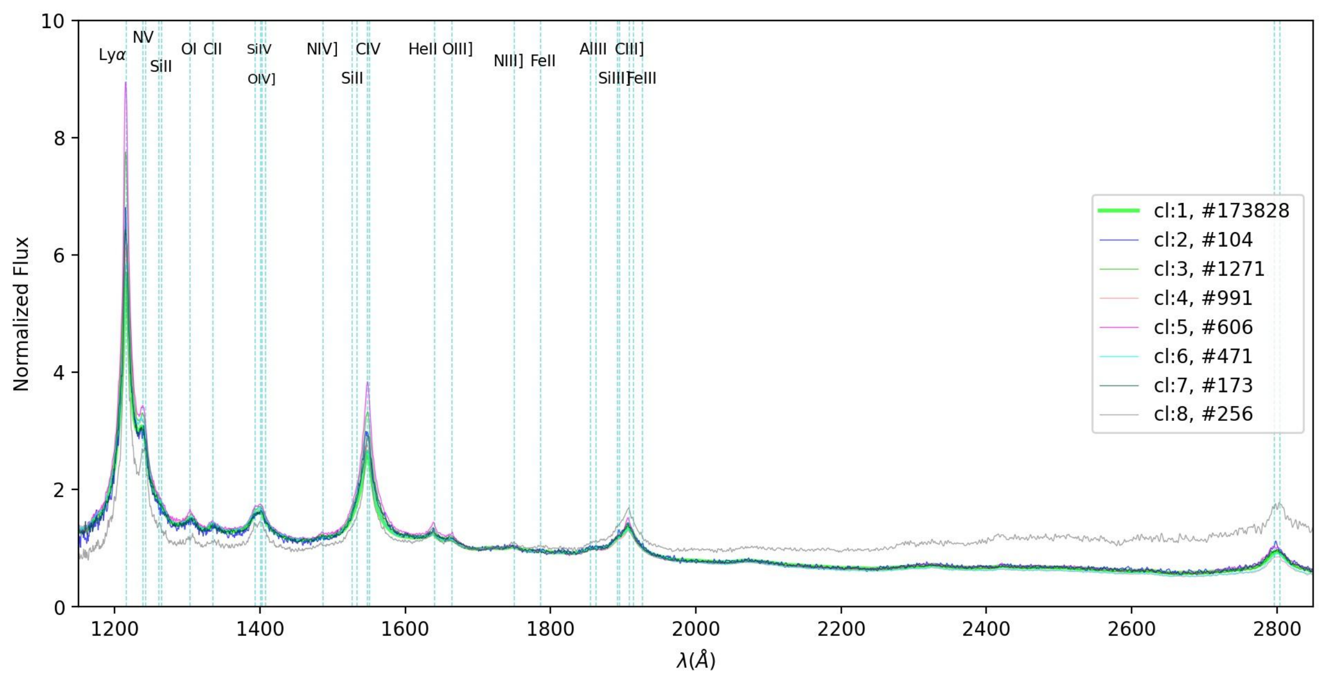
# Evaluation with Isolation Forest scores



Distinction in iForest scores distribution of clusters show they are more outlier than the biggest cluster.

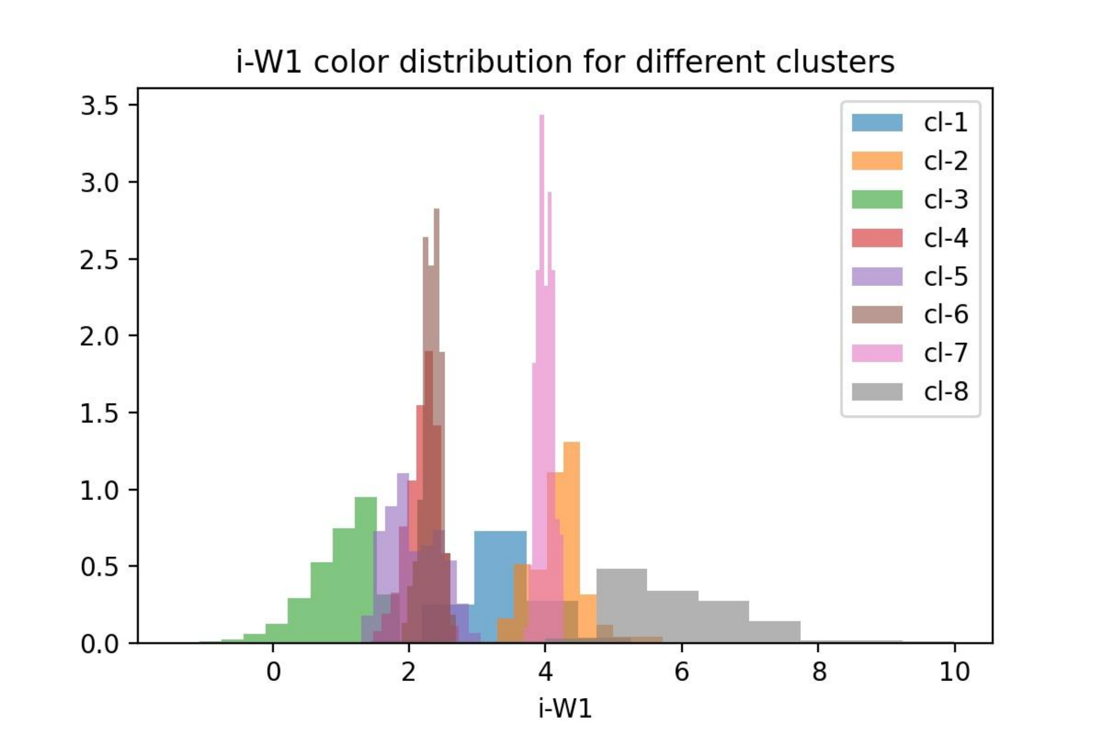
# Evaluation with median spectra

- Median spectra for each cluster of quasars
- Cluster 8 has the most distinct spectrum
- Cluster 8 is very isolated in tSNE map



# Evaluation with color distribution

- i-W1 color distributions for different clusters
- This validates the distinct color behavior of quasars
- Cluster 8 has very red quasars
- Cluster 3 has very blue quasars



# Conclusion

1. We found some clusters of outlier quasars
2. Among the outlier cluster, a cluster consisted of 256 objects was more interesting.
3. Our most anomalous cluster:
    a. Has the highest Isolation Forest anomaly scores
    b. Has the most deviated median spectrum
    c. Has the reddest color among other clusters