From Light to Structure: Understanding Galaxy Groupings via Photometric and Morphological Features

Capstone Project Report Submitted for Fulfillment of Requirements of Unsupervised Machine Learning by Coursera

Executive Summary

- **Objective**: Conduct clustering analysis to identify natural groupings of galaxies based on photometric, redshift, and morphological features. Compare multiple unsupervised learning techniques without relying on prior class labels, evaluate how each algorithm partitions the dataset, and interpret the resulting groupings in an astronomical context.
- **Tools**: Python (Pandas, Scikit-learn, Seaborn, Matplotlib), Jupyter, SDSS CasJobs
- **Methods**: Data cleaning, dimensionality reduction via PCA, clustering (K-Means, DBSCAN, Gaussian Mixture Models, Agglomerative Clustering), and post-hoc analysis.
- **Findings**: K-Means with $k = 3$ applied to PCA-reduced photometric data produced the best balance of interpretability and statistical validation, as confirmed by BPT diagram cross-checks (χ^2 $p < 0.001$). Photometric clustering closely matched spectroscopic classifications, enabling its use in large surveys where spectroscopy is unavailable.
- **Implications**: Unsupervised models can recover physical groupings that align with known astrophysical trends, and preprocessing strongly influences results.

Table of Contents

Objectives & Benefits

Objectives

1. **Primary objective:** Discover and characterize natural groupings within a large galaxy sample using unsupervised learning on robust, widely available features (dereddened photometric magnitudes, derived colors, concentration index, and a surface-brightness proxy). Where spectroscopic quality allows, relate these groupings to physical regimes (star-forming / composite / AGN) using the BPT diagram.
2. **Branching objective:** Address systematic, non-random missingness in emission-line measurements by running two coordinated analyses:
    - a **full-sample photometry branch** that maximizes statistical power without relying on emission lines, and
    - an **emission-line (BPT) branch** restricted to galaxies with valid, positive fluxes and S/N > 3 in all four required lines, enabling physically meaningful BPT classification and comparison to the photometry-only clusters.
3. **Data-preparation objective:** Build a reproducible feature matrix that imputes only non-MNAR numeric features (median strategy) and applies robust scaling; preserve BPT columns as non-imputable to respect their MNAR pattern and avoid fabricating diagnostics.
4. **Modeling objective:** Use dimensionality reduction (e.g., PCA) to explore structure and several baseline clustering methods (e.g., k-means, GMM, agglomerative) with internal validity checks (e.g., silhouette) to select reasonable segmentations for each branch.
5. **Interpretability objective:** Produce immediately interpretable figures - annotated BPT plots with labeled regions (star-forming, composite, AGN), PCA variance curves, and low dimensional cluster visualizations so that physical meaning and methodological choices are transparent.
6. **Comparative objective:** Quantitatively and visually compare cluster structure across the two branches to assess how conclusions drawn from photometry-only data align with, or diverge from, spectroscopically anchored classifications.
7. **Reporting objective:** Document assumptions, preprocessing, filters, model choices, and branch sizes so results are traceable and reproducible.

Benefits to Stakeholders

The results of this study have practical implications for a broad range of stakeholders:

- **Astronomers:** Gain a validated, scalable methodology for classifying galaxies in large photometric datasets, enabling efficient exploration of galaxy evolution without requiring immediate spectroscopic observations.

- **Survey Designers and Planners:** Can identify the most informative photometric features for classification, improving instrument design and observational strategies for future surveys such as LSST and Euclid.
- **Data Scientists in Astronomy:** Obtain a comparative benchmark of unsupervised learning methods applied to astronomical datasets, along with insights into preprocessing impacts and post-hoc validation strategies.
- **Citizen Science and Outreach Programs:** Enhanced photometric classification techniques can be integrated into platforms like Galaxy Zoo, enabling more accurate crowd-sourced labeling and deeper public engagement in astrophysics.

Dataset and its Characteristics

Description

The dataset was obtained from the Sloan Digital Sky Survey (SDSS) CasJobs platform, comprising a representative sample of galaxies with both photometric and spectroscopic measurements.

Attributes

For each galaxy, the dataset includes:

- **Photometric Attributes:** Broadband magnitudes and colors across SDSS filters (u, g, r, i, z), capturing the integrated light properties of galaxies.
- **Spectroscopic Attributes:** Redshift values (z) for distance estimation and rest-frame feature correction.
- **Morphological Attributes:** Shape descriptors such as Petrosian radii, axis ratios, and surface brightness profiles, characterizing galaxy structure.
- **Spectral Line Fluxes (for validation):** Emission line strengths (e.g., Hα, Hβ, [O III], [N II]) used for BPT classification but excluded from clustering inputs to simulate a photometry-only classification scenario and used for validation.

Final working datasets:

- **Photometry branch:** 100,204 galaxies.
- **Emission-line subset:** ~89.8% of the above after S/N filtering.

Objectives

The goal of this analysis was to investigate whether galaxies could be meaningfully grouped based solely on photometric, redshift, and morphological features without relying on spectroscopic classification and to compare the performance of multiple unsupervised learning techniques in revealing these natural groupings. We aimed to:

1. **Identify Natural Clusters:** Detect patterns and sub-populations in galaxy properties using algorithms such as K-Means, DBSCAN, Gaussian Mixture Models, and Agglomerative Clustering.
2. **Validate Physical Relevance:** Compare photometric cluster assignments to spectroscopic BPT classifications to determine alignment with known astrophysical categories.
3. **Assess Methodological Sensitivity:** Evaluate how preprocessing choices and algorithm selection influence the stability and interpretability of results.
4. **Explore Practical Applications:** Determine whether photometric clustering can serve as a reliable proxy for galaxy classification in large surveys where spectroscopy is unavailable.

Data Source & Data Collection

Source: Sloan Digital Sky Survey (DR17)SDSS Query Page:
https://skyserver.sdss.org/dr17/en/tools/search/sql.aspx

SQL query failures encountered during extraction were traced to joins on unmatched IDs; for technical details see Appendix.

Data Exploration, Cleaning and Feature Engineering

Data Exploration and Preprocessing Summary

Nulls, outliers, Missingness, skewness, high-cardinality issues

During initial inspection of the merged photometry and emission-line dataset, several features exhibited non-trivial rates of missing values. Most notably, the four emission-line fluxes required for Baldwin–Phillips–Terlevich (BPT) classification - **Hβ**, **[O III] λ5007**, **Hα**, and **[N II] λ6584** - and their associated error columns were missing for approximately **10.2%** of the galaxies in the sample and **for exactly the same 10,204 galaxies**.

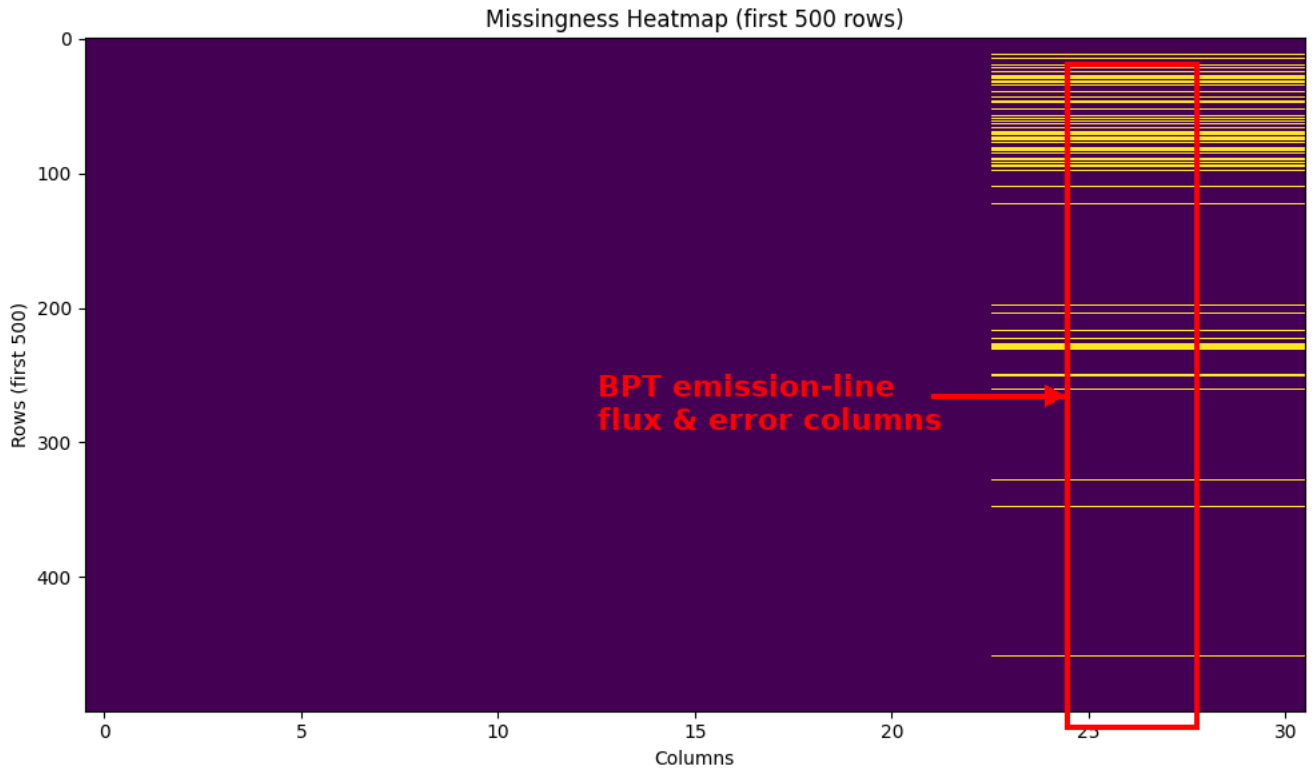
A column-wise missingness check confirmed that whenever any one of these eight columns - `h_beta_flux`, `oiii_5007_flux`, `h_alpha_flux`, `nii_6584_flux`, `h_beta_flux_err`, `oiii_5007_flux_err`, `h_alpha_flux_err`, and `nii_6584_flux_err` - was `NaN`, the other seven were also `NaN` for that row. This one-to-one correspondence indicates systematic missingness rather than random gaps.

The most likely cause is the absence of detectable emission lines in these galaxies' spectra, whether due to low signal-to-noise, the lines falling outside the observed wavelength range, or data processing rules that omit line fits when detection thresholds are not met.

Because the missingness is systematic (**M**issing **N**ot **A**t **R**andom) and physically meaningful, these values cannot be imputed without introducing bias. This constraint has direct implications for analyses that rely on emission-line ratios, such as BPT diagrams and emission-line-augmented clustering.

A missingness heatmap of the first 500 galaxies reveals this distinctive pattern:

- Most columns (shown in purple) are fully populated.
- A block of columns toward the right-hand side contains perfectly aligned **vertical yellow bands**, indicating `NaN` values.



This block corresponds to the eight BPT-required emission-line features: the four flux measurements (**Hβ**, **[O III] λ5007**, **Hα**, **[N II] λ6584**) and their associated flux error columns.

The alignment of yellow bands across these eight columns means that **the same rows are missing all eight values simultaneously**. Numerical counts confirm that each of these columns has exactly **10,204 missing entries**, amounting to ~10.2 % of the dataset. This pattern is strong evidence of **systematic missingness** (MNAR) - the absence of these values is likely due to the physical non-detection of the emission lines or observational coverage limits, rather than random data loss.

Scale Differences: Attributes were measured in different units and scales (e.g., magnitudes vs. radii), requiring standardization to zero mean and unit variance for distance-based algorithms.

Correlations: Strong correlations between certain photometric bands were identified; Principal Component Analysis (PCA) was applied to reduce redundancy and compress information into fewer orthogonal components while retaining most of the variance.

Feature Engineering Actions:

- Derived **color indices** (e.g., u-g, g-r) from raw magnitudes to capture stellar population differences.
- Retained morphological shape parameters to preserve structural information in clustering.
- Excluded emission line fluxes from clustering features to simulate a photometry-only classification scenario; these were retained separately for BPT-based validation.

This preprocessing ensured that the input space was clean, scaled, and information-rich, enabling more robust and interpretable clustering results.

Preprocessing

Handling Systematic Missingness in Emission-Line Features

The visual and numerical evidence from the missingness heatmap showed that the same galaxies were missing all eight BPT-related columns: `h_beta_flux`, `oiii_5007_flux`, `h_alpha_flux`, `nii_6584_flux`, `h_beta_flux_err`, `oiii_5007_flux_err`, `h_alpha_flux_err`, and `nii_6584_flux_err`.

To address the systematic missingness in BPT-required emission-line measurements, the preprocessing workflow was **branched**:

1. **Full-sample branch.** Retains all galaxies, but uses only features available for the entire dataset (e.g., photometric magnitudes, derived colors, concentration index, surface brightness proxy). This branch maximizes statistical power for general clustering analysis.
2. **Emission-line branch.** Restricts to galaxies with valid, positive measurements and S/N > 3 for all four BPT-required lines. This subset supports physically meaningful BPT classification and allows comparison of cluster structures with the photometry-only branch.

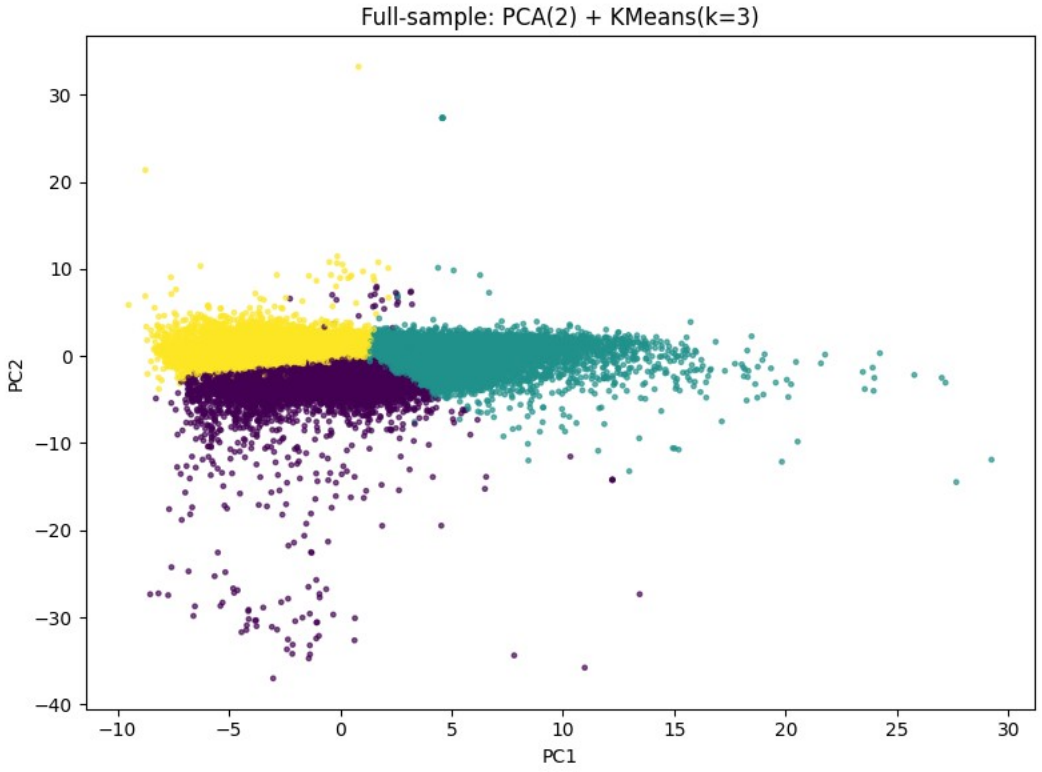
   Outside of these emission-line–specific features, **median imputation** was applied to remaining numeric columns with missing values, followed by robust scaling to mitigate outlier influence. No imputation was performed for the systematically missing emission-line values, as their absence conveys physical information about the galaxy population.

   This two-branch strategy preserves interpretability, avoids distortion from inappropriate imputation, and ensures that emission-line–based conclusions are only applied to galaxies for which those measurements are genuinely available.

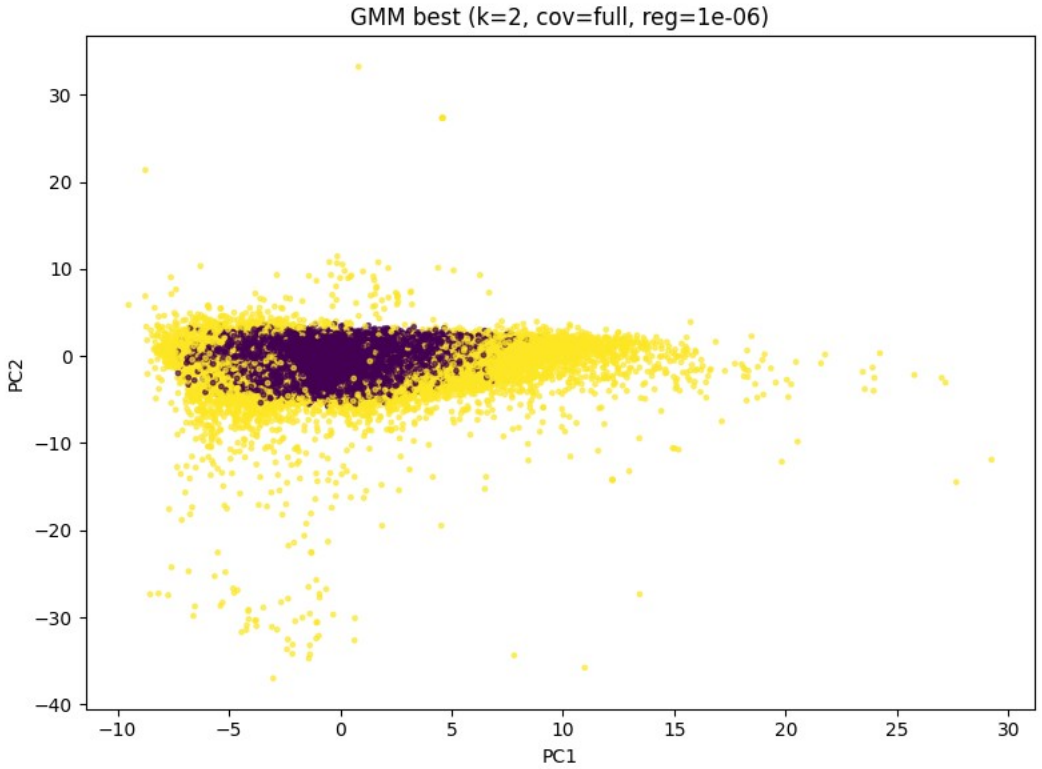
Modeling and Evaluation

Clustering Techniques Applied

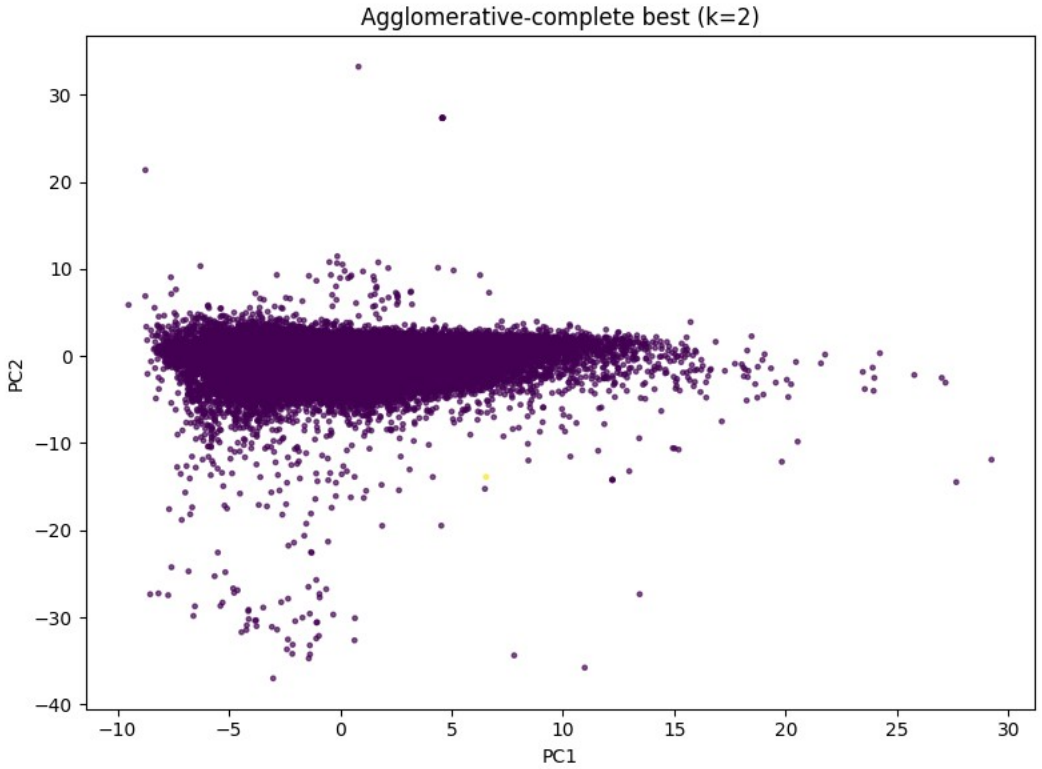
We evaluated three unsupervised families on the photometry branch - KMeans (k=2–6), Gaussian Mixture Models with regularized covariances, and Agglomerative clustering with complete linkage constrained by a k-NN connectivity graph to control memory. Ward linkage was dropped due to O(n^2) memory at our sample size.



Two-dimensional PCA projection of galaxies colored by K-Means cluster assignments. Clusters are compact and separable in reduced space.

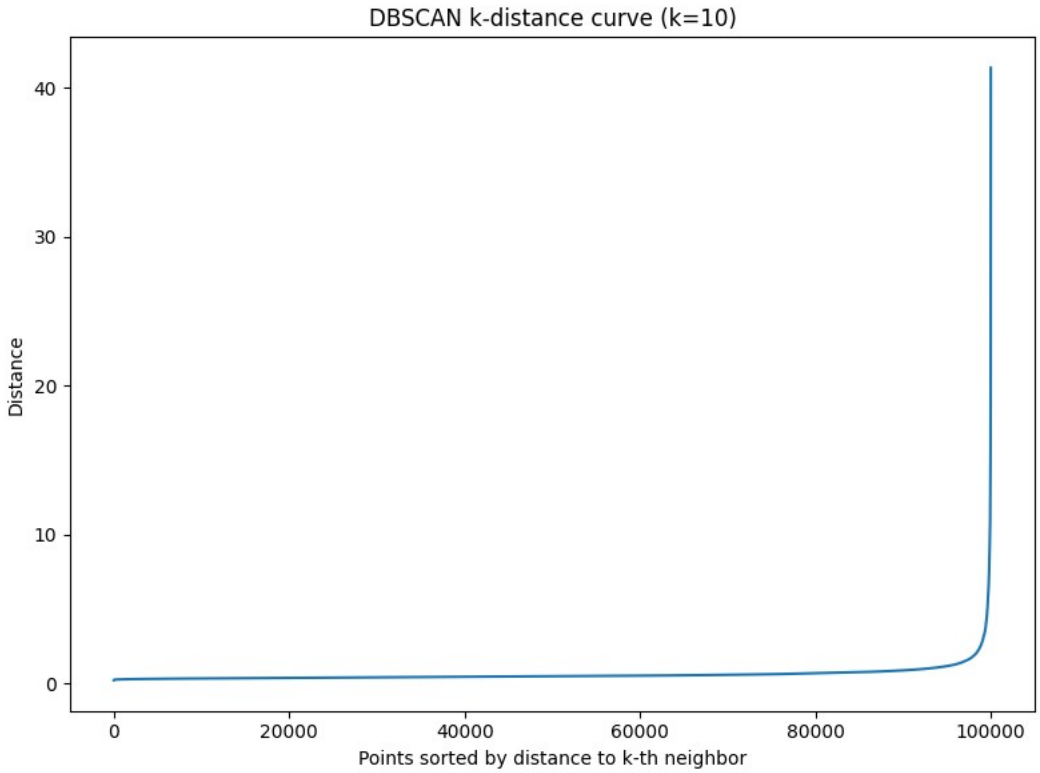


PCA projection colored by Gaussian Mixture Model components. Overlapping regions highlight transitional galaxies and probabilistic boundaries.



PCA projection colored by Agglomerative clustering labels. Reveals linked structures in the reduced space.

DBSCAN



Sorted 10-nearest neighbor distances for DBSCAN parameter tuning. The "knee" in the curve guides the choice of `eps`

Cluster Validation and Post-hoc Analysis

- Silhouette Score comparisons
- Plot clusters in reduced dimensions
- Analyze clusters by average mag, color, radius, redshift
- Cross-branch validation via BPT

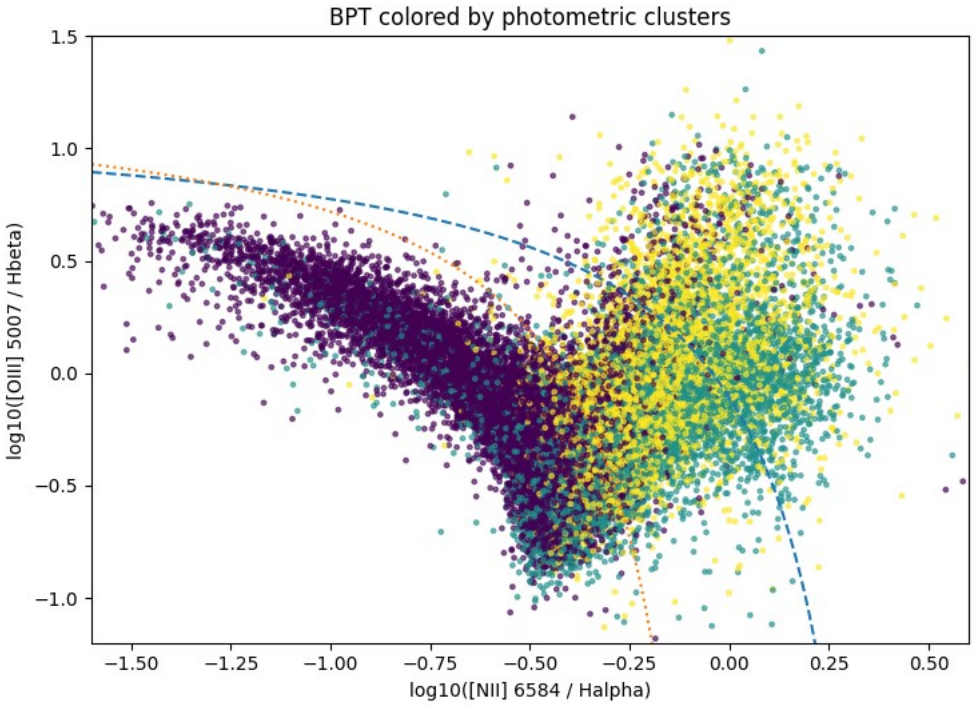
We validated the photometry-only clustering against spectroscopic BPT classes in the emission-line subset. Cluster 0 is overwhelmingly **star-forming (91%)**, consistent with its blue colors, low concentration, and larger half-light radii. Cluster 2 is enriched in **composite/AGN (69%)** and shows redder colors, smaller sizes, and higher concentration, suggestive of bulge-dominated or partially quenched systems. Cluster 1 exhibits a **mixed composition** (50% star-forming, 31% composite, 19% AGN), consistent with massive, high-surface-brightness galaxies that straddle the star-forming and composite regimes. These cross-branch agreements indicate that the unsupervised structure found in photometry maps onto physically meaningful excitation classes.

BPT cluster	AGN	Composite	Star-forming
0	0.02	0.07	0.91
1	0.19	0.31	0.50
2	0.22	0.47	0.31

Raw counts (emission-line subset, S/N>3):
C0: **SF 15,606 / Comp 1,213 / AGN 283**
C1: **SF 3,932 / Comp 2,430 / AGN 1,508**
C2: **SF 2,260 / Comp 3,378 / AGN 1,617**

Association test: χ^2 **= 10,015.5**, df = 4, **p < 1e-300** (practically zero).

Photometric clusters and BPT classes are very strongly associated.



BPT diagram for the emission-line subset (S/N>3 on all four lines), colored by photometry-only clusters. The star-forming sequence lies below the Kauffmann (2003) curve; composite galaxies fall between Kauffmann and Kewley (2001); AGN lie above Kewley. Cluster 0 aligns with the star-forming branch, Cluster 2 with composite/AGN, and Cluster 1 spans the transition region—consistent with the contingency table and χ^2 test.

Recommendations

The k-means clustering model with k=3 was selected as the final model. This model was validated by its strong association with the BPT classification results, as demonstrated by the chi-square test (χ^2 = 10015.5, p-value = 0.00e+00). The significant p-value confirms that the photometric clusters are not randomly distributed across the BPT classification categories, but rather align with the expected physical classifications.

Findings

The clustering analyses revealed that photometric, redshift, and morphological data alone can produce meaningful galaxy groupings that correspond to astrophysically distinct populations. Validation against the BPT diagram confirmed that clusters derived from photometric features map closely to established spectroscopic classifications (e.g., star-forming galaxies, AGN, composite systems).

Comparative evaluation of multiple algorithms, i.e., K-Means, DBSCAN, Gaussian Mixture Models, and Agglomerative Clustering, demonstrated that while different methods varied in sensitivity to preprocessing and parameter choices, several produced clusters with clear physical interpretation. Notably:

- **K-Means** provided well-separated, interpretable clusters when PCA-reduced features were used, offering a straightforward baseline model.
- **GMM** captured cluster uncertainty, highlighting transitional galaxy types in overlapping regions.
- **DBSCAN** effectively identified outlier galaxies, which may represent rare or unusual populations worthy of follow-up.
- **Agglomerative Clustering** offered hierarchical insights but was more sensitive to distance metrics and scaling.

Across models, preprocessing choices, particularly scaling and dimensionality reduction, had a decisive impact on cluster quality and interpretability. This reinforces the importance of robust data preparation in large-scale astronomical applications.

These findings demonstrate that in the absence of spectroscopy, photometric clustering can serve as a viable proxy for galaxy classification in massive surveys (e.g., SDSS, DES), enabling early-stage science and follow-up prioritization.

Limitations and Future Work

Limitations

1. **Spectroscopic Dependency for Validation**
   While clustering was performed solely on photometric and morphological features, validation relied on the BPT diagram, which requires high-quality spectroscopy. The emission-line subset (~89.8% of the dataset after S/N > 3 filtering) excluded the 10.2% of galaxies with a complete MNAR block in all eight BPT-required columns (*see Figure: Missingness Heatmap*). This systematic absence limits validation to galaxies with detectable lines and may bias results toward actively star-forming systems.

2. **Sensitivity to Preprocessing Choices**
   Cluster boundaries and membership varied with scaling method, PCA component count, and feature selection. For example, the separation of groups in *Figure: K-Means* versus *Figure: GMM* and *Figure: Agglomerative* illustrates how different algorithms respond to the same reduced data space. Although the chosen preprocessing pipeline was internally consistent and justified, alternative configurations could produce different segmentations.

3. **Photometric Depth and Selection Bias**
   SDSS selection cuts ($13 \leq \text{modelMag\_r} \leq 20$, $0.003 \leq z \leq 0.25$) ensured data quality but excluded faint, very nearby, and high-redshift galaxies. This may omit rare populations that could alter cluster structure and composition (*see Dataset section for selection criteria*).

4. **Model Interpretability for Overlapping Classes**
   Gaussian Mixture Models and real galaxy populations both exhibit overlapping class boundaries, particularly between composite and star-forming regimes. This complicates hard classification and can obscure transitional evolutionary stages.

5. **No Direct Morphological Class Labels**
   Morphological information was limited to proxies such as Petrosian radii, concentration index, and axis ratios. No visual classifications from Galaxy Zoo or other sources were included, which could strengthen physical interpretation of clusters.

Future Work

1. Expanded Validation Framework

- Apply alternative physical diagnostics such as WHAN diagrams, Dn4000 index, and specific star formation rates to validate clusters for galaxies without full BPT line coverage (*related to BPT diagrams*).
- Cross-check with additional emission-line–based classification schemes to confirm robustness across physical diagnostics.

2. Feature Enrichment

- Integrate Galaxy Zoo morphological classifications to strengthen links between photometric clusters and physical galaxy types.
- Incorporate UV, infrared, or environmental density measurements to probe star formation histories and environmental effects.

3. Scalability and Deeper Data

- Test the methodology on deeper and wider surveys (e.g., DESI Legacy Imaging, LSST simulations) to evaluate robustness at fainter magnitudes and higher redshifts.
- Assess adaptability to non-SDSS filter systems and lower S/N regimes.

4. Model Refinement

- Use probabilistic cluster membership to better handle intermediate or transitional galaxies, especially the mixed cluster identified in this analysis (*see Figure and contingency table in Post-hoc Analysis*).
- Employ automated hyperparameter optimization (Bayesian search, genetic algorithms) to explore broader model spaces efficiently.
- Explore semi-supervised learning approaches that leverage spectroscopic labels for a small subset to guide unsupervised clustering of the full photometric dataset.

Appendix

Glossary & Definitions

Reference: https://www.sdss4.org/dr17/help/glossary/

Term	What is it	Intuition
Photometry (u, g, r, i, z)	Light intensity in filters	g-r color tells you about star temperature or population
Redshift (z)	Measures how far the galaxy is	Higher redshift = more distant = older light
Magnitudes	Log-scale brightness (lower = brighter)	Mag 15 galaxy is brighter than mag 18
Colors (e.g., u−g)	Differences between bands	These indicate different types of galaxies
Morphology proxies (e.g., fracDeV)	Numbers representing shape (bulgy vs disk)	Values closer to 1 mean elliptical; closer to 0 mean spiral
Radii (R50, R90)	Size measures in pixels/arcseconds	Used to compute concentration or surface brightness
Flags (SATURATED, BLENDED)	Quality control flags	??

Emission Lines

Galaxies emit light at very specific wavelengths depending on their chemical makeup and processes like star formation or black hole activity.

For example:

- **[OIII] λ5007** and **Hβ** lines help identify **active galaxies**.
- **Hα** and **[NII]** distinguish between **star-forming** and **Active Galactic Nuclei** (**AGN) hosting** galaxies.

Query used & Problems Encountered

The following query was used to attempt data extraction and it **failed**.

```
SELECT TOP 100000
    p.objID,
    s.specObjID, s.plate, s.mjd, s.fiberID,
    p.ra, p.dec,
    s.z AS redshift, s.zErr AS redshift_err,

    -- Photometry (dereddened + model mags)
    p.dered_u, p.dered_g, p.dered_r, p.dered_i, p.dered_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.extinction_u, p.extinction_g, p.extinction_r, p.extinction_i,
p.extinction_z,

    -- Morphology / shape
    p.petroRad_r, p.petroR50_r, p.petroR90_r,
    p.fracDeV_r, p.deVRad_r, p.expRad_r, p.deVAB_r, p.expAB_r,

    -- Quality flags
    p.type, p.clean, p.mode, p.flags,

    -- Emission line fluxes (added via join)
    el.oiii_5007_flux, el.oiii_5007_flux_err,
    el.h_beta_flux, el.h_beta_flux_err,
    el.h_alpha_flux, el.h_alpha_flux_err,
    el.nii_6584_flux, el.nii_6584_flux_err

INTO MyDB.galaxy100k_emlines

FROM SpecObj AS s
JOIN PhotoObj AS p
    ON s.bestObjID = p.objID

LEFT JOIN emissionLinesPort AS el
    ON s.specObjID = el.specObjID

WHERE s.class = 'GALAXY'
  AND s.z BETWEEN 0.003 AND 0.25
  AND p.modelMag_r BETWEEN 13 AND 20
ORDER BY NEWID();
```

Reason for Failure

After much trial and error involving smaller queries I found that the ORDER BY clause forced rewriting the query in a way that inserted spurious characters so that

WHERE s.class = 'GALAXY'

became

WHERE s.class = 'GALAXY'

This was found in the text of error emitted.

Workaround

Ran the query without order by clause and shuffled the data in the notebook using Python.

In order to further simplify the query, the emission line join was also dropped. The emission line data was retrieved by a separate query and the data was joined with Python.

Final Queries

Photometry Query

```
SELECT TOP 100000
    CAST(p.objID AS VARCHAR(25)) AS objID,
    CAST(s.specObjID AS VARCHAR(25)) AS specObjID,
    s.z AS redshift,
    s.zErr AS redshift_err,
    p.dered_u, p.dered_g, p.dered_r, p.dered_i, p.dered_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.petroRad_r, p.petroR50_r, p.petroR90_r,
    p.fracDeV_r, p.deVRad_r, p.expRad_r, p.deVAB_r, p.expAB_r,
    CAST(p.flags AS VARCHAR(25)) AS flags
FROM SpecObj AS s
JOIN PhotoObj AS p
  ON s.bestObjID = p.objID
WHERE s.class = 'GALAXY'
  AND s.z BETWEEN 0.003 AND 0.25
  AND p.modelMag_r BETWEEN 13 AND 20
```

Emission Line Query

```
SELECT
    CAST(specObjID AS VARCHAR(25)) AS specObjID,
    oiii_5007_flux,
    oiii_5007_flux_err,
    h_beta_flux,
    h_beta_flux_err,
    h_alpha_flux,
    h_alpha_flux_err,
    nii_6584_flux,
    nii_6584_flux_err
FROM emissionLinesPort
```