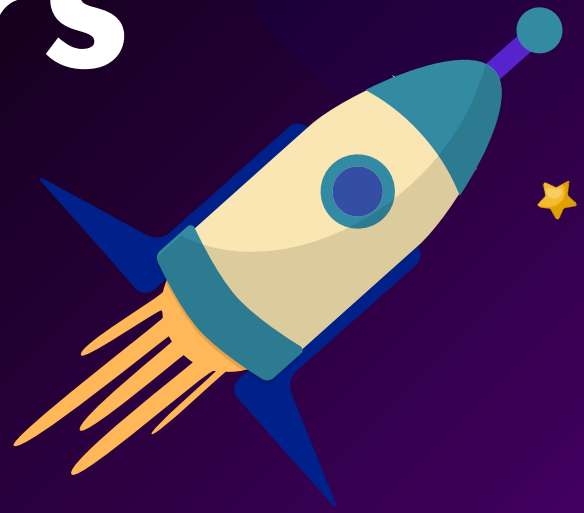
Predicting

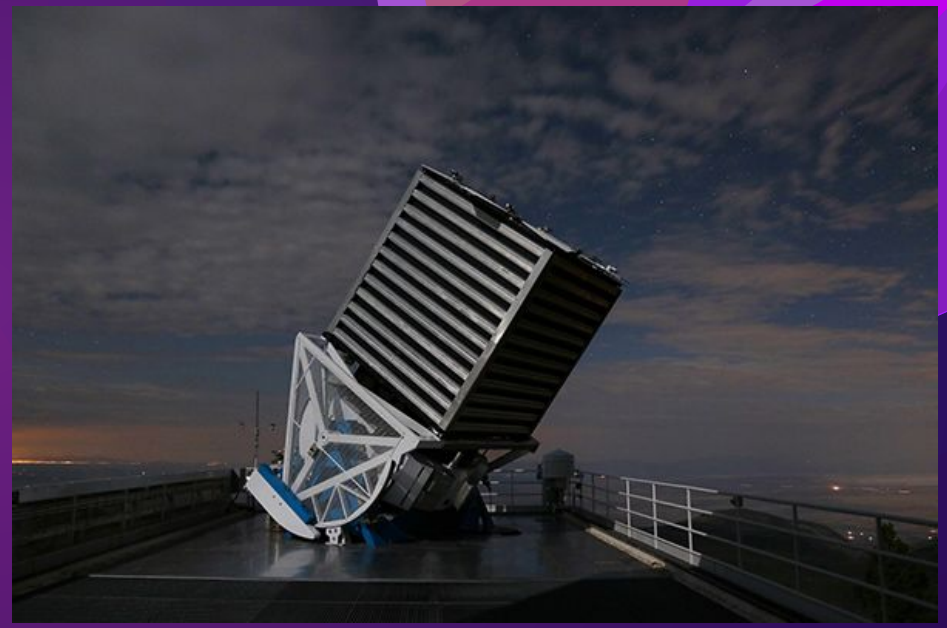
Stars, Galaxies and Quasars

Team 8: Aditi Nankar, Lily Saltonstall,
Ian O'Connor, Shiki Pawan



Sloan Digital Sky Survey

The Sloan Digital Sky Survey is a facility that maps, tracks, and photographs the cosmos. They are pioneers of panoptic spectroscopy, collecting data gathered from their many ongoing cosmic surveys. They dataset we used for our project was found on Kaggle.com and features over 10,000 rows of data from nearby and distant stars, galaxies, and quasars!

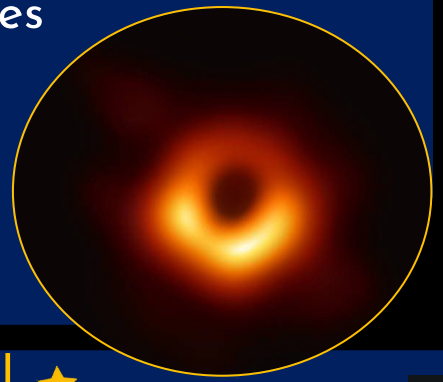


The Sloan Digital Sky Survey Telescope

Image Credit: Patrick Gaulme, Data Release 17 | SDSS

Galaxies

Galaxies are massive celestial bodies containing multitudes of stars and solar systems within them. Many large galaxies have a supermassive black hole at their center.



Stars

Fueled by radiation, they emit light across the electromagnetic spectrum. Stars with light in the ultraviolet or blue wavelengths are hotter and stars with red wavelengths are cooler.

Quasars

The most luminous objects in the cosmos, giving off light across the entire spectrum. These beams of raw energy are emitted by supermassive black holes at the core of galaxies. Particles are trapped by the powerful magnetic field and are violently ejected from the magnetic poles.

Fun Fact! ★

The quasar located nearest to Earth is called Markarian 231 and is located in the constellation Ursa Major (the big dipper).

Cygnus a is one of the first quasars ever discovered! Pictured here

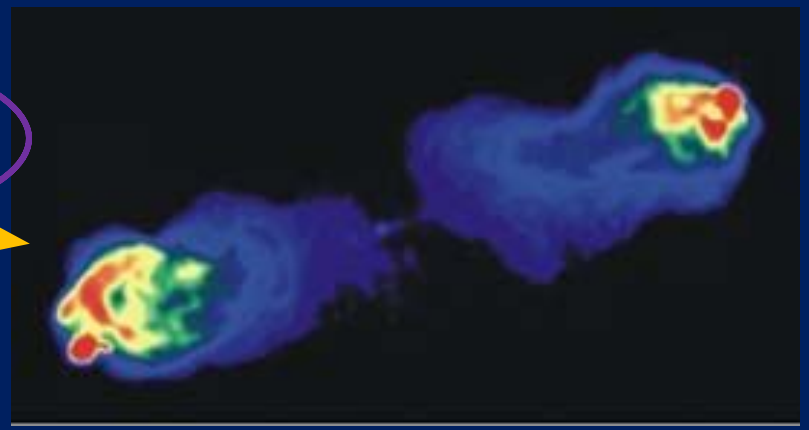


Image Credit: https://voyages.sdss.org/expeditions/universe/quasars/early-radio-astronomy/

Gas and dust whip around the black hole in a swirling vortex of terror known as an accretion disk. The gas is heated up to temperatures that are millions of degrees and cause the quasar to emit thermal radiation that is so bright it can be seen across the light spectrum.

The Data Used (and what it means)

<u>SDSS photometric system:</u>
These are optical filters used in a telescope known as passbands. The filters ensure high efficiency for faint object detection and essentially cover the entire accessible optical wavelength range.

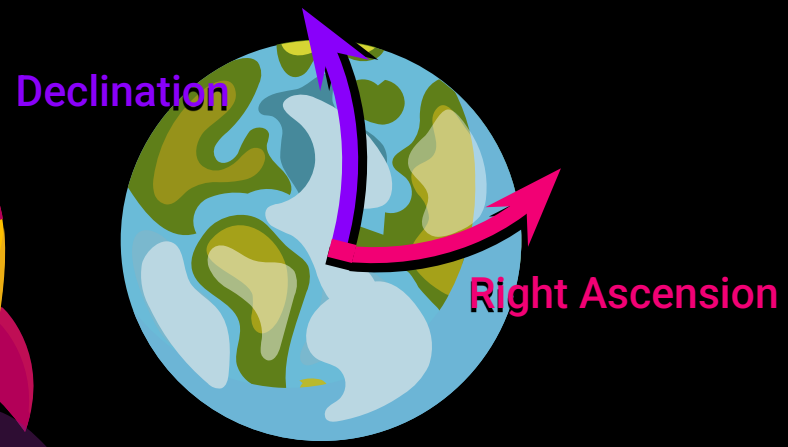
u = Ultraviolet/Blue Wavelengths
g = Green Wavelengths
r = Red Wavelengths
i = Near Infrared
z = Infrared

The SDSS's Spectrograph (seen from the side)
Image Credit: https://skyserver.sdss.org/dr1/en/proj/basic/color/fromstars.asp

The Data Used (continued)

Coordinate System

ra = Right Ascension (longitude)
dec = Declination (latitude)



Redshift

The stretching of wavelength and the shifting of light towards the red end of the light spectrum as celestial objects move away from Earth.

Redshifting can be compared to the Doppler Effect which is the same concept but with sound and on a smaller scale.

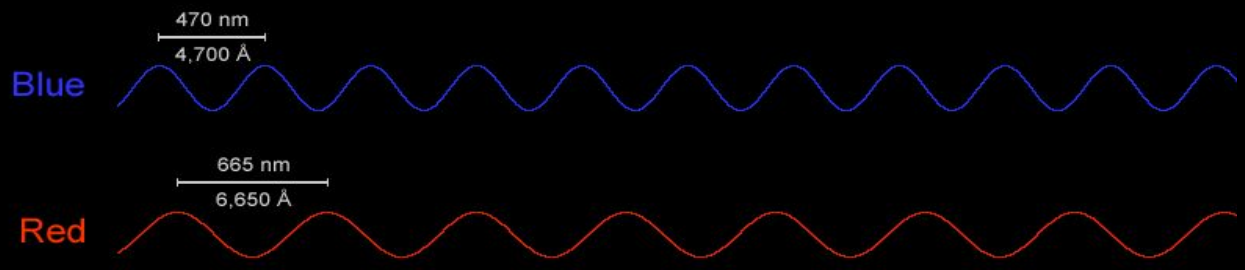
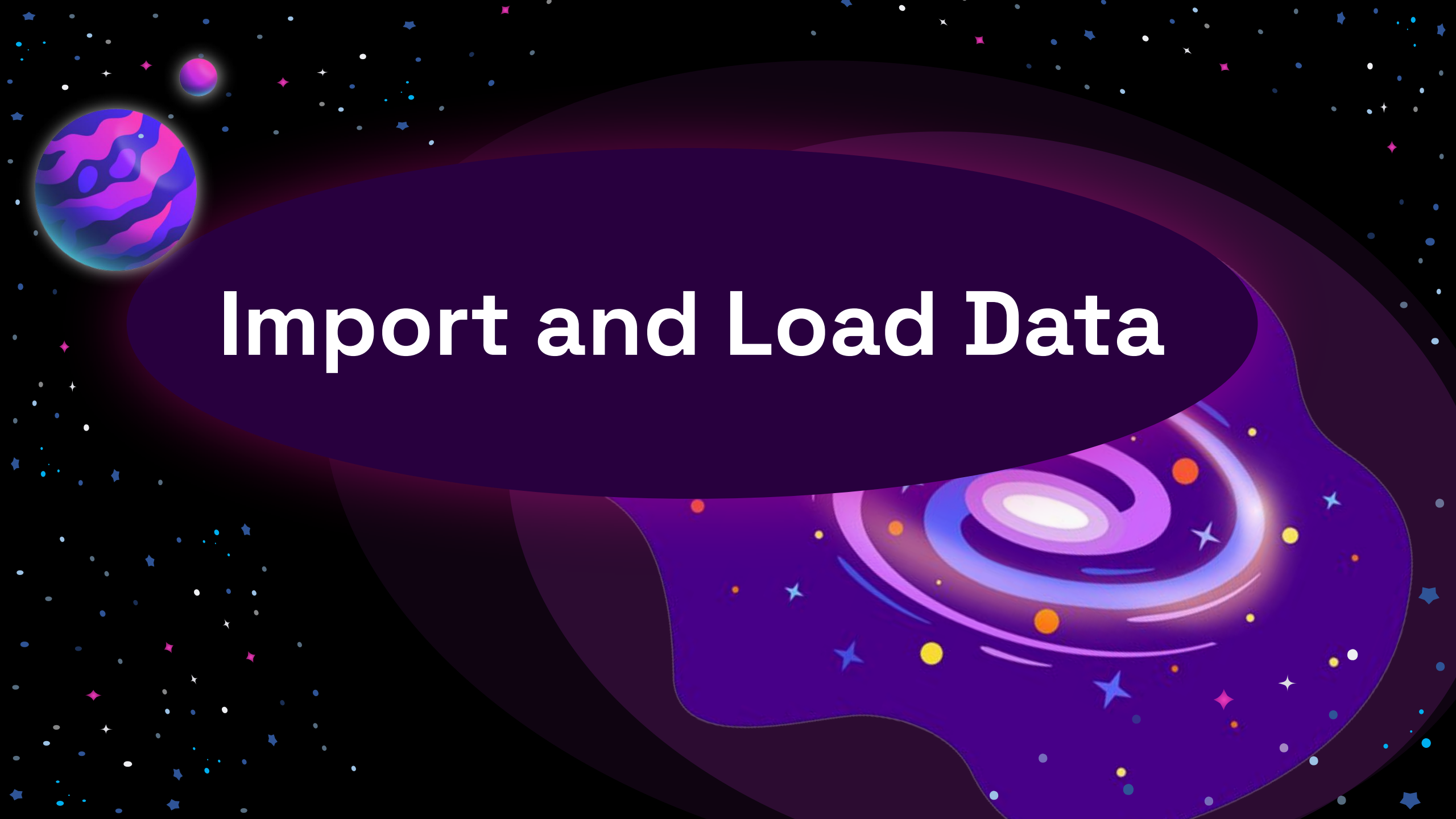
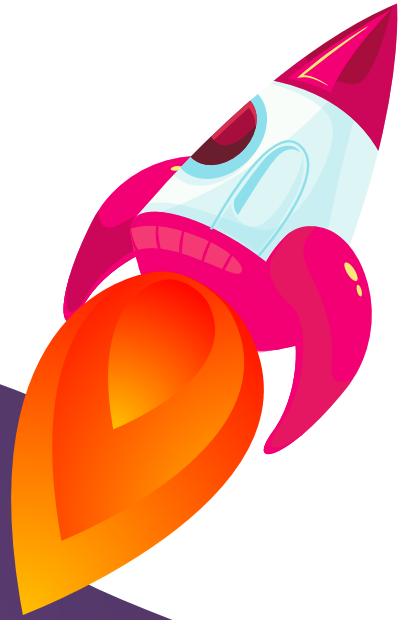


Image Source: https://scied.ucar.edu/image/wavelength-blue-and-red-light-image


Import and Load Data

Imported Libraries



- matplotlib
- seaborn
- pandas
- numpy
- plotly
- sqlalchemy
- sklearn
- tqdm

SQL Database Integration

1. Create

Created our 'StarsGalaxiesQuasars' database and schema code

2. Populate

Imported our Sloan Sky Survey CSV data from Kaggle via Postgres

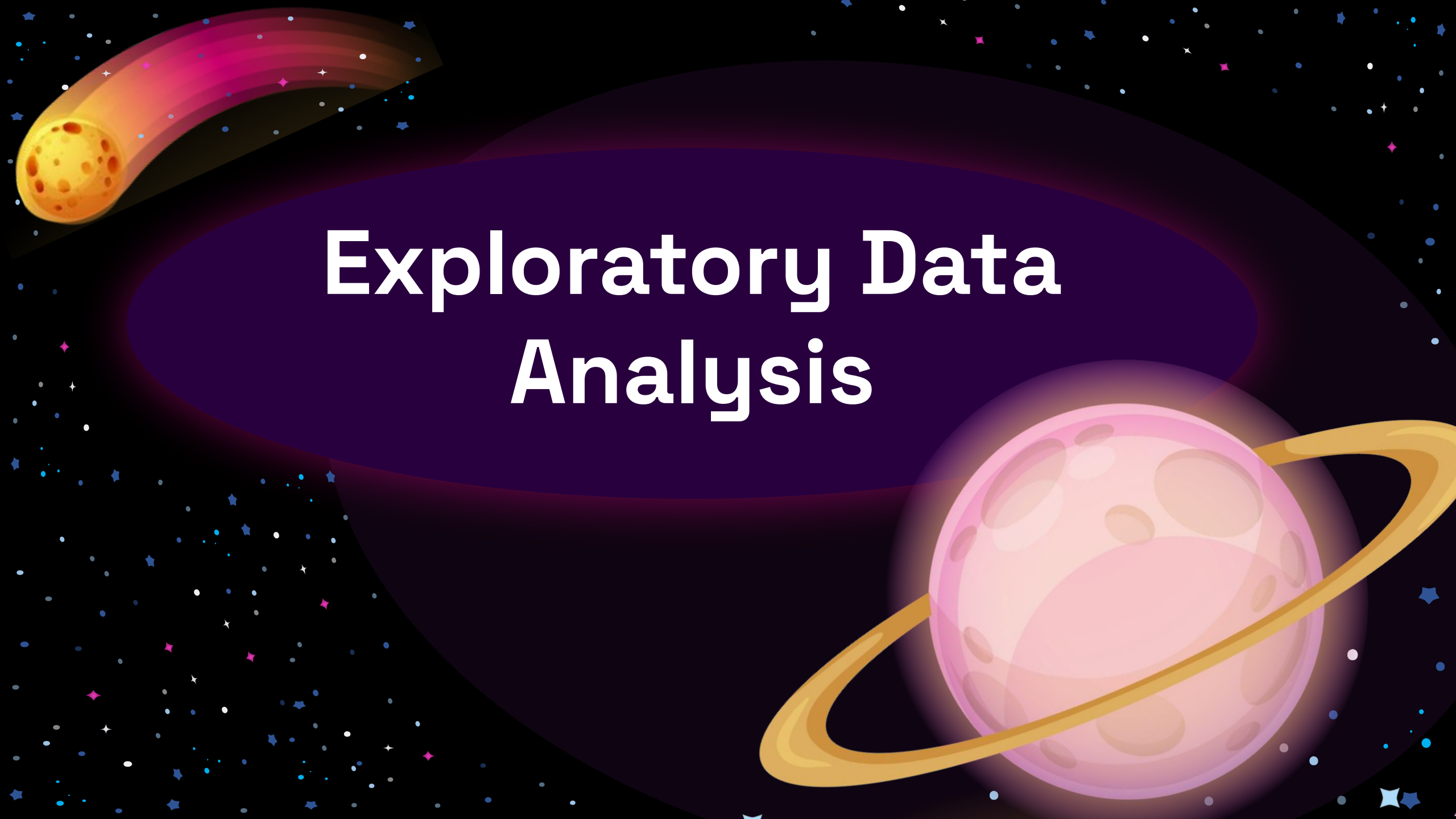
3. Connect

Established a database connection in our JupyterLab Notebook using SQL Alchemy

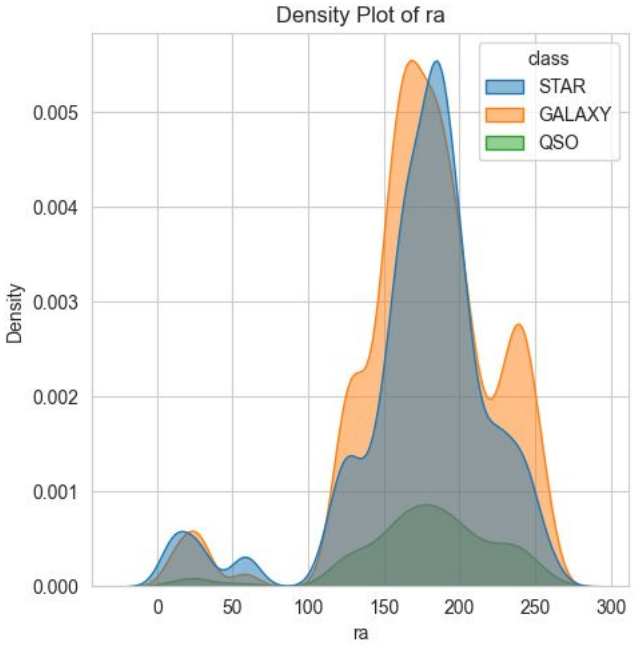
4. Load

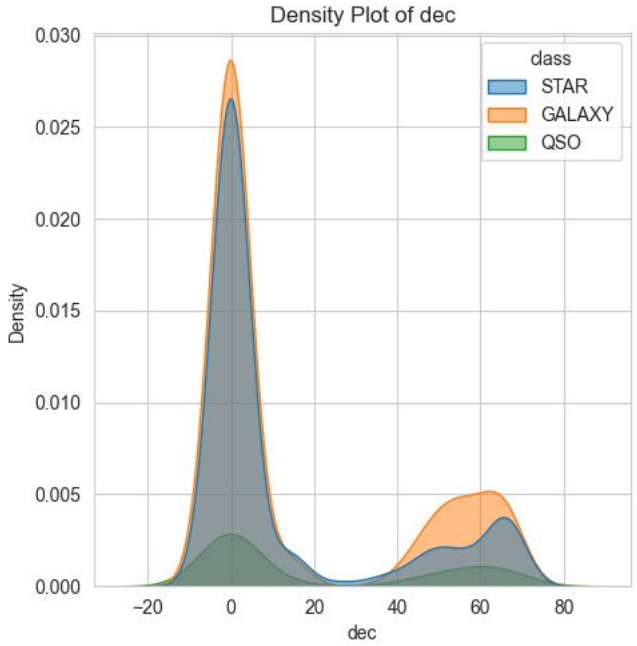
Loaded the results of our SELECT queries into a Pandas dataframe for analysis

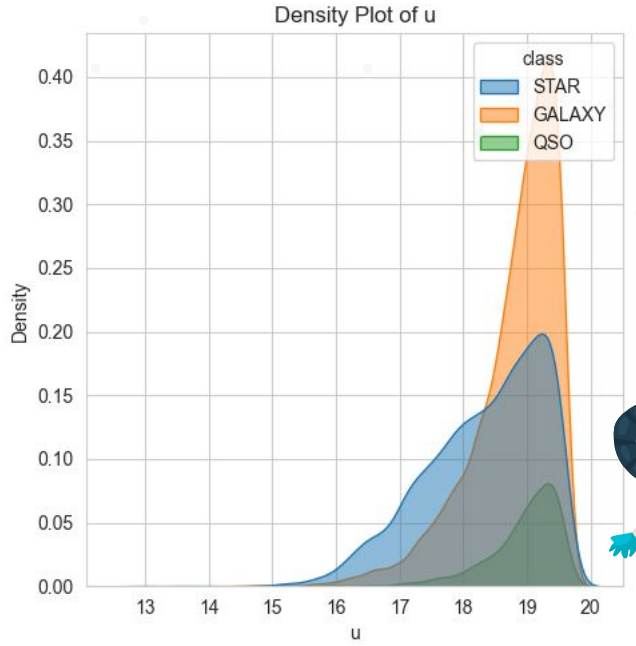
Exploratory Data Analysis

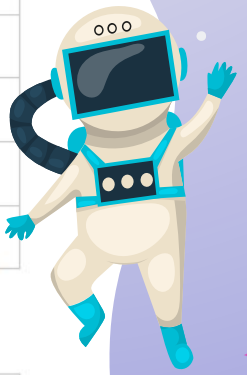


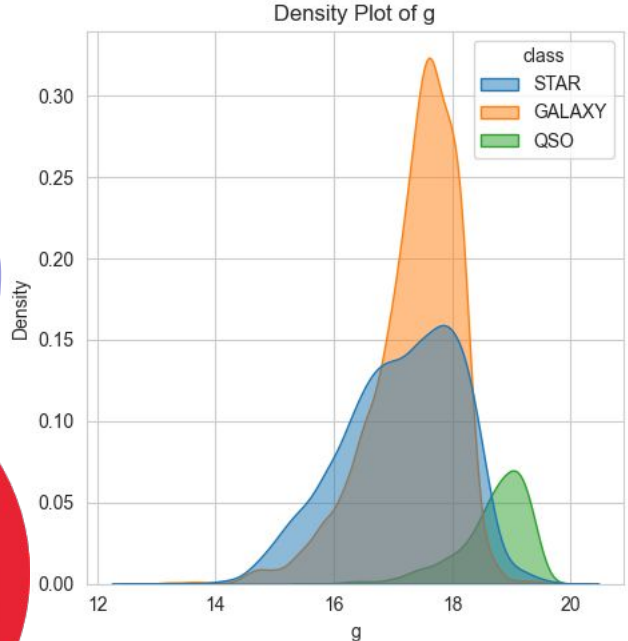
Density Plot Visual

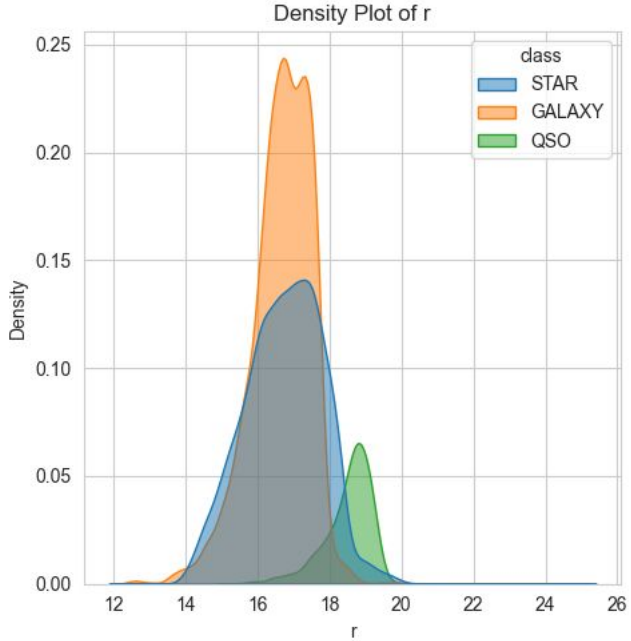


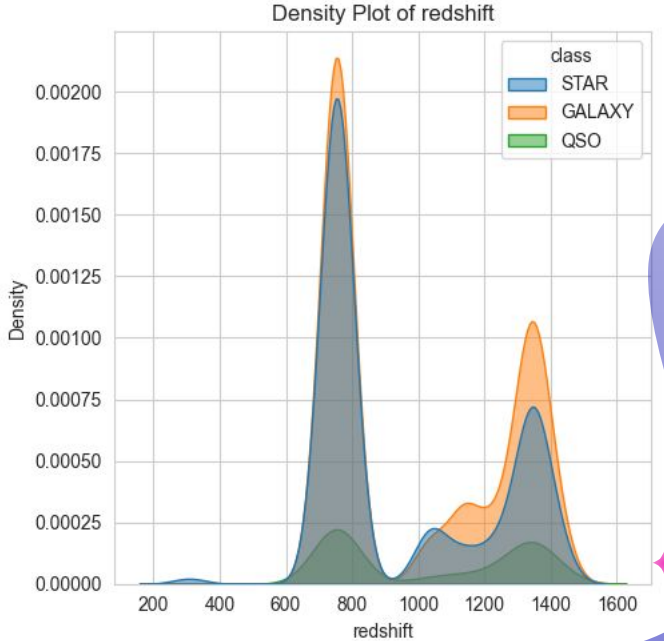








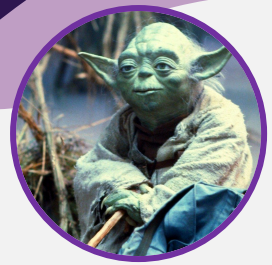




Findings

Correlation

Redshift is the most distinguishing feature – stars have low redshift, galaxies have medium redshift, and quasars have high redshift



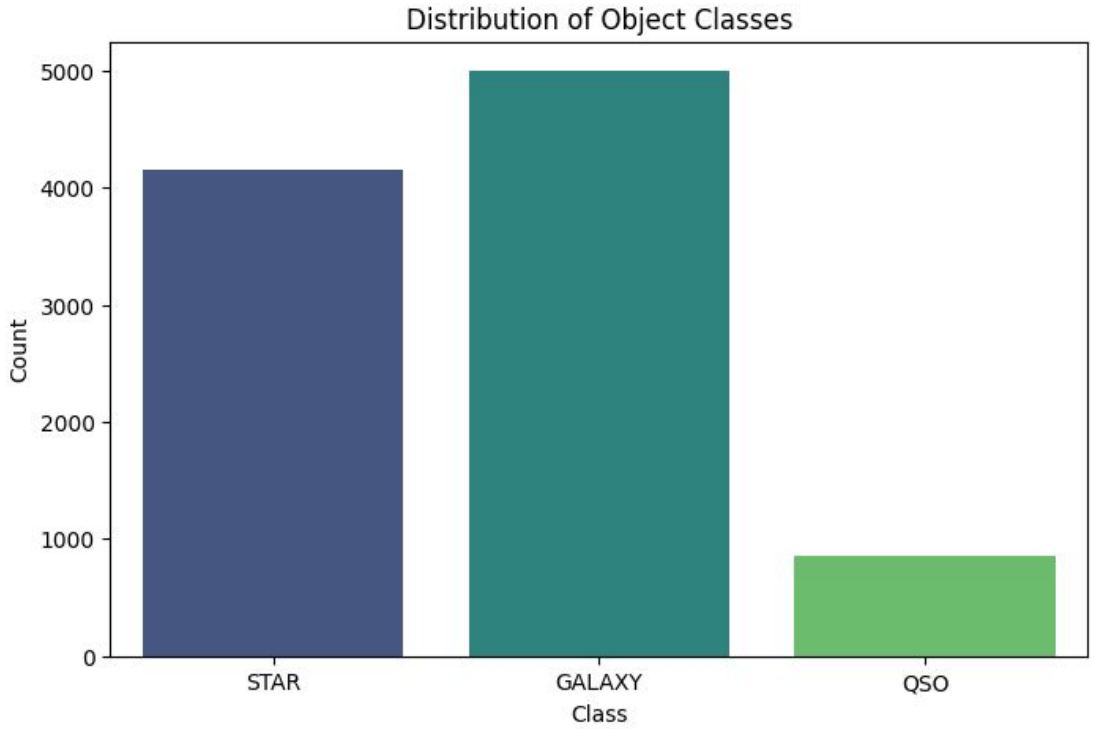
Takeaway

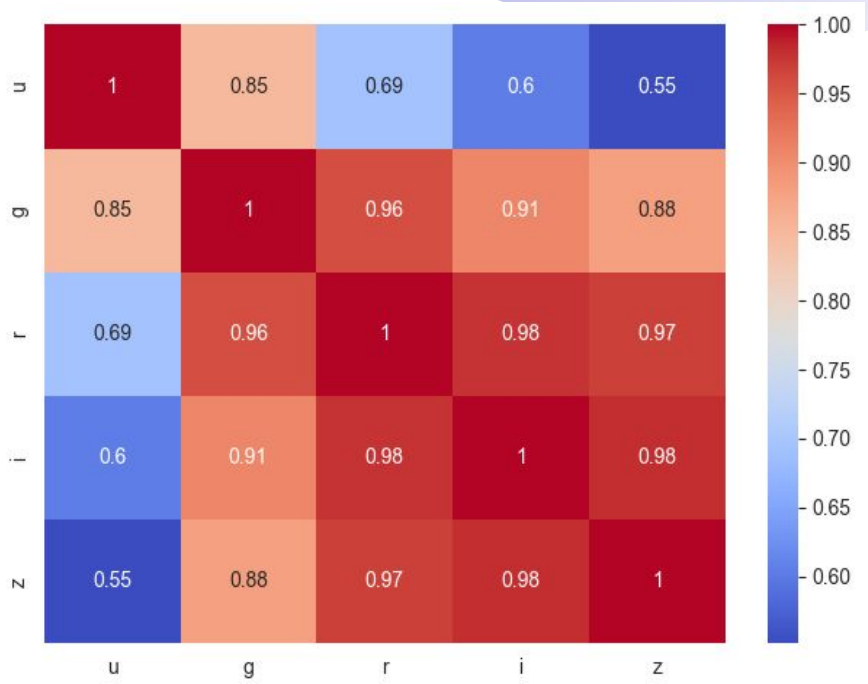
Because of its direct correlation with our target variables, we decided to remove the redshift column from our features to help mitigate data leakage

Fun fact

There's a giant space blob of alcohol on a distant cloud of gas called, Sagittarius B2. There's enough alcohol to make 400 trillion pints of beer

Exploratory Data Analysis Visuals





Findings

Correlation

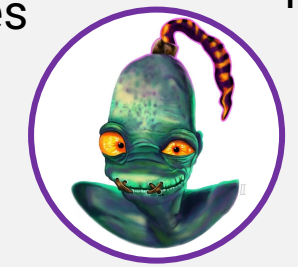
Even without the redshift, our correlation heatmap still shows heavy correlation among our key data features

Takeaway

Data imbalance and heavy feature correlation may cause issues with our training and analysis

Imbalance

Our count plot shows us that there are significantly more stars and galaxies in our target (Y) data compared to quasars



Fun Fact

Footprints left on the Moon won't disappear as there is no wind

Data Pre-Processing

Filtering Our Dataset

Unused Columns:

objid = Object Identifier
specobjid = Object Identifier
run = Run Number
rerun = Rerun Number
camcol = Camera Column
plate = Plate Number
field = Field Number
fiberid = Fiber ID
ra = Right Ascension
dec = Declination
mjd = Modified Julian Date
redshift = Object's Redshift Value

Used Columns:

class = Class (galaxy, star, quasar)

u = Ultraviolet/Blue Wavelengths
g = Green Wavelengths
r = Red Wavelengths
i = Near Infrared
z = Infrared

Key Features Used

U

Ultraviolet/blue

G

Green

R

Red

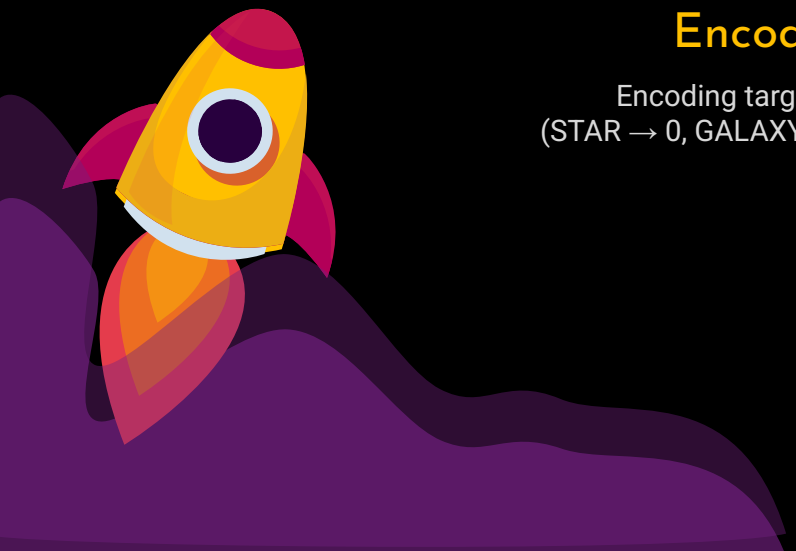
I

Near infrared

Z

Infrared

Other Data Preprocessing Steps



Encoding

Encoding target variable
(STAR → 0, GALAXY → 1, QSO → 2)

Splitting

Splitting dataset into training (80%)
and testing (20%) sets

Scaling

Normalizing values using
StandardScaler

Machine Learning Models

Random Forest

Results:

	precision	recall	f1-score	support
star	0.92	0.95	0.94	832
galaxy	0.96	0.93	0.94	999
qso	0.89	0.89	0.89	169
accuracy			0.94	2000
macro avg	0.92	0.93	0.92	2000
weighted avg	0.94	0.94	0.94	2000

SVM

(Support Vector)

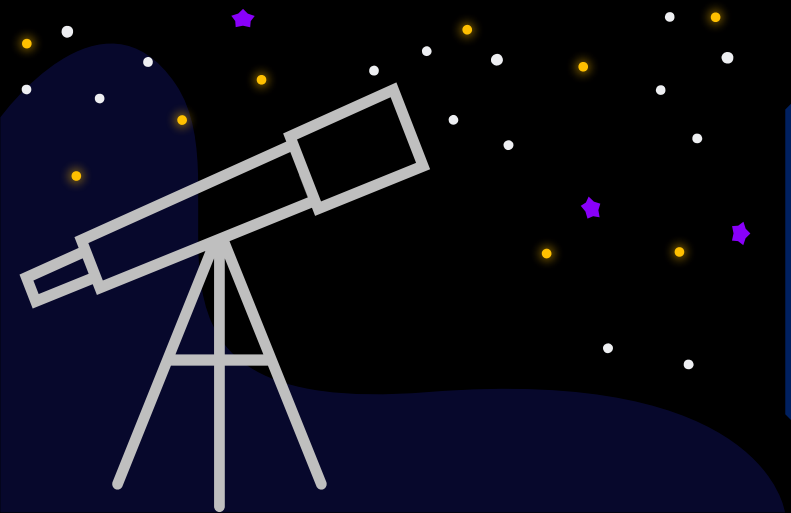
Results:

	precision	recall	f1-score	support
star	0.92	0.97	0.94	832
galaxy	0.95	0.94	0.95	999
qso	0.96	0.79	0.86	169
accuracy			0.94	2000
macro avg	0.94	0.90	0.92	2000
weighted avg	0.94	0.94	0.94	2000

Logistic Regression

Results:

	precision	recall	f1-score	support
star	0.91	0.96	0.93	832
galaxy	0.96	0.92	0.94	999
qso	0.93	0.90	0.92	169
accuracy			0.94	2000
macro avg	0.93	0.93	0.93	2000
weighted avg	0.94	0.94	0.94	2000



Gradient Boosting

Results:

	precision	recall	f1-score	support
star	0.84	0.90	0.87	832
galaxy	0.92	0.87	0.89	999
qso	0.90	0.85	0.87	169
accuracy			0.88	2000
macro avg	0.89	0.87	0.88	2000
weighted avg	0.88	0.88	0.88	2000

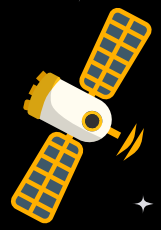
Check for Data Leakage

Step 1

Examine and thoroughly research all data features used and determine their impact and/or significance

Step 2

Gradually modify select data features and test repetitively for any changes in accuracy



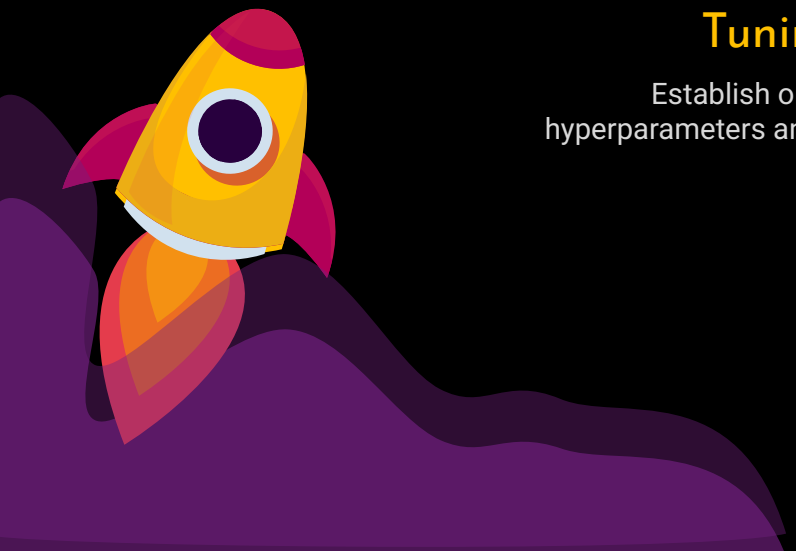
Gradient Boosting Model

Gradient boosting is a prediction model similar to Random Forest commonly used in sales forecasting, stock price prediction, fraud detection, and financial modeling, Gradient boosting models are commonly used to win Kaggle competitions due to their accuracy.

Boosting improves model performance by correcting the mistakes of previous learners with new ones. It uses weak learners, like decision stumps, which perform slightly better than random guessing. These weak learners are sequentially added, focusing on correcting errors at each step.

A key difference between the gradient boosting model and the random forest model is overfitting. While adding more trees in random forests doesn’t cause overfitting, in gradient boosting, too many weak learners can lead to overfitting the data. This makes careful hyperparameter tuning even more essential when using gradient boosting models.

Steps for GB Model Optimization



Tuning

Establish optimum hyperparameters and train on them

Visuals

Observe feature importance and confusion matrix visual aids

PCA

Improve efficiency and performance with data compression

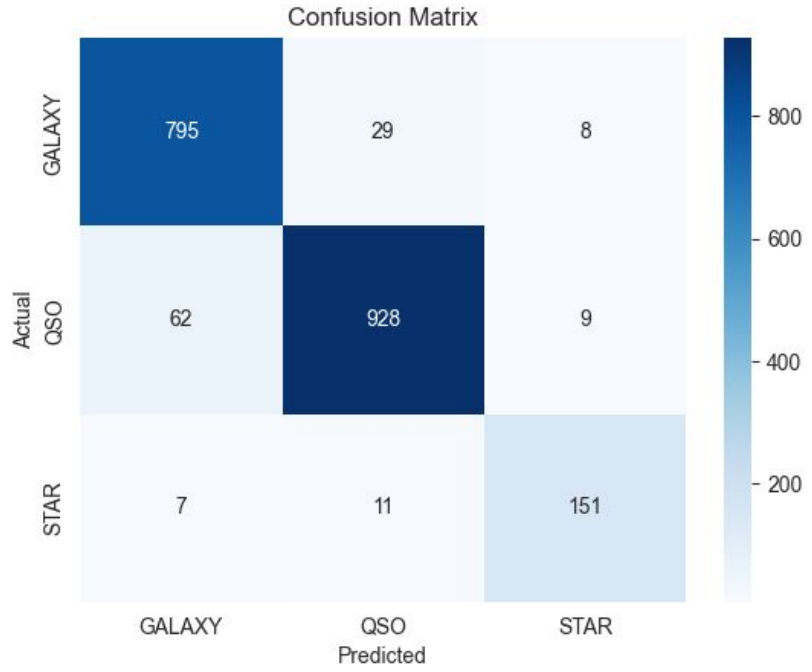
GB Hyperparameter Tuning

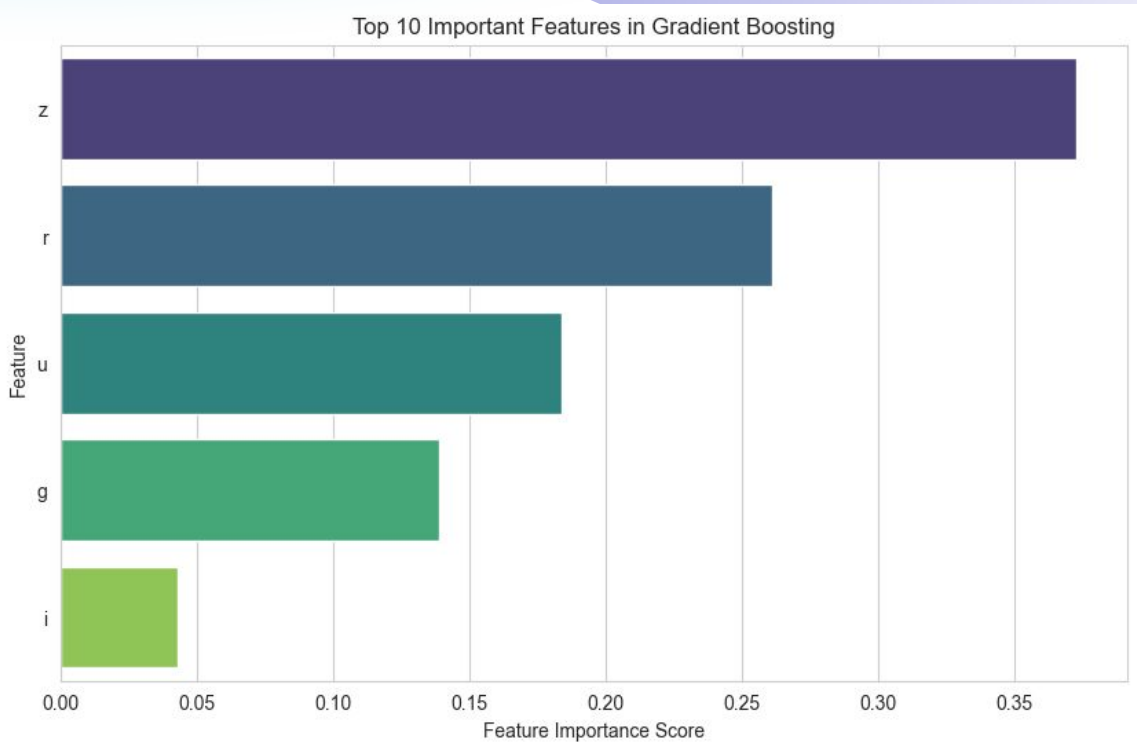
Used **RandomizedSearchCV** to search for best hyperparameters.

Best Parameters Found:

- n_estimators=300
- learning_rate=0.2
- max_depth=7
- min_samples_leaf=2
- subsample=0.9

Tuned GB Parameter Visuals:





Findings

Balance

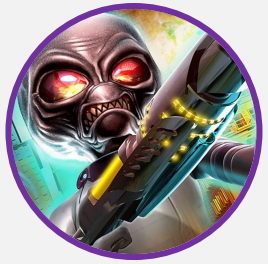
Overall, the feature importance is decently balanced, however, prediction strength relies most heavily on z (Infrared), possibly due to its ability to capture distant objects like Quasars

Takeaway

With high classification accuracy and decently balanced prediction strength of the feature (X) data, this model is well optimized

Accuracy

Classification accuracy is high, however, misclassification occurs most between Quasars and Galaxies due to their spectral similarities



Fun Fact

Star Wars was a documentary

Hyperparameter Tuning Final Results:

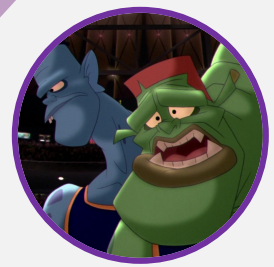
Optimized Gradient Boosting Model Report:

	precision	recall	f1-score	support
star	0.92	0.96	0.94	832
galaxy	0.96	0.93	0.94	999
qso	0.90	0.89	0.90	169
accuracy			0.94	2000
macro avg	0.93	0.93	0.93	2000
weighted avg	0.94	0.94	0.94	2000

Findings

Accuracy

Our accuracy rates increased across all margins with hyperparameter tuning



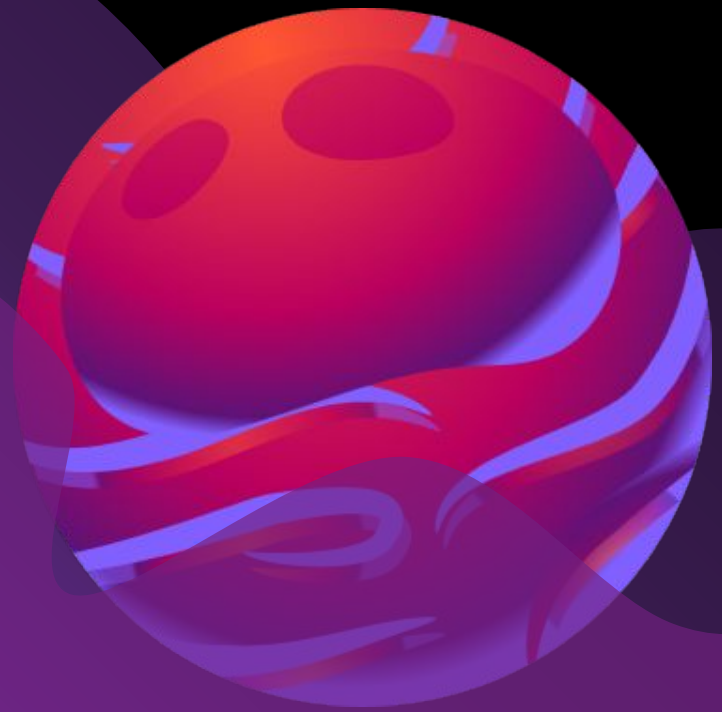
Takeaway

With parameter tuning, we were able to increase our model's total predictive performance by over 6%

Fun Fact

1 tsp of a neutron star weighs the same as the human population

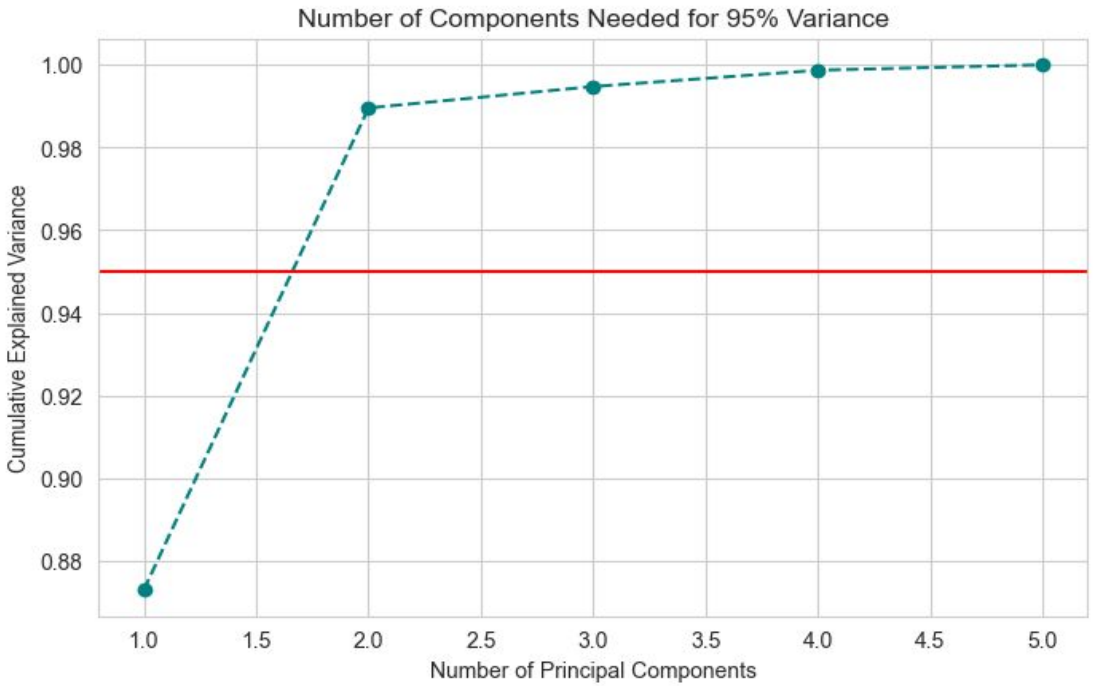
Optimization with PCA

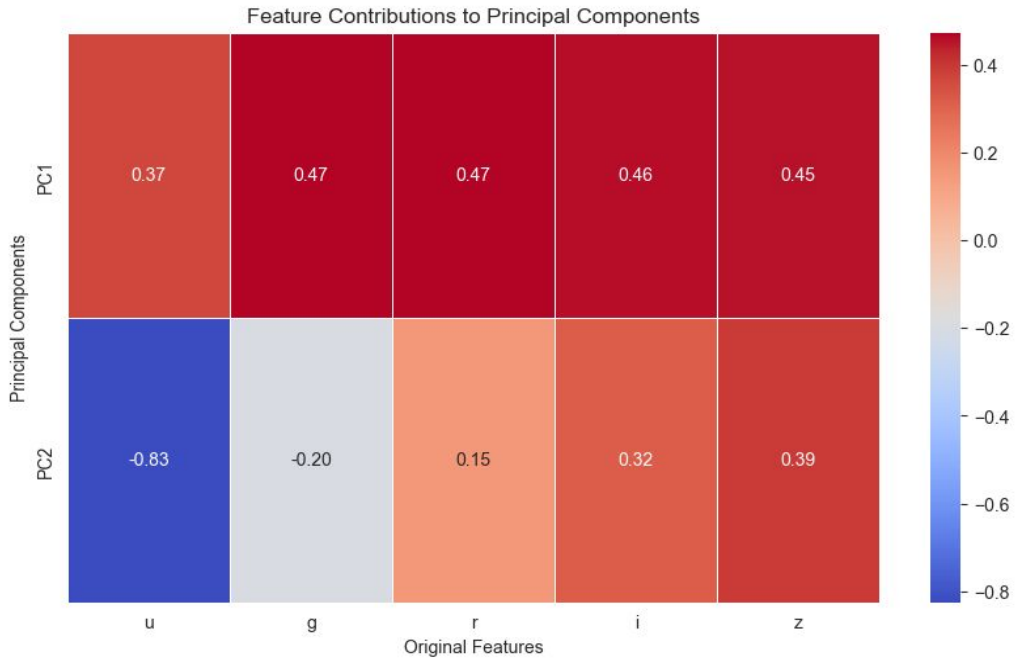


Advantages of Dimensionality Reduction:

- Computational Efficiency
- Model Performance
- Overfitting Prevention
- Better Visualization

PCA Visual Aids





PCA Results:

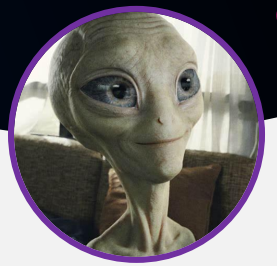
Gradient Boosting Model with PCA Classification Report:

	precision	recall	f1-score	support
star	0.71	0.70	0.71	832
galaxy	0.78	0.79	0.78	999
qso	0.81	0.82	0.81	169
accuracy			0.75	2000
macro avg	0.77	0.77	0.77	2000
weighted avg	0.75	0.75	0.75	2000

Findings

Accuracy

Despite high correlation, PCA data compression actually decreases our accuracy by nearly 20%



Takeaway

Using PCA for optimization is not a great fit for our dataset despite encouraging indications

Correlation

The projection plot shows that our data is highly correlated as only two principal components are needed to explain 95% of the data variance

Fun Fact

Uranus' blue glow is due to the gases in its atmosphere

Conclusion

Using the Gradient Boosting Classifier we were able to train our machine learning model to distinguish between stars, galaxies, and quasars, achieving very high accuracy through systematic hyperparameter optimization.

Overall, our project pipeline integrates database extraction, model evaluation, and effective data visualization strategies. This work contributes to the automated classification of astronomical objects and demonstrates the broader application of machine learning in astrophysics.


Thank
You!


We are now opening the space gates for questions—fire away!