DSE 615 Data Science in Practice Final Project Report

Spectroscopic classification of Photometric data from the Sloan Digital Sky Survey Data Release 17 using Statistical techniques in Machine Learning and Deep Learning

By

Rita Abani 19244

Department of Electrical Engineering and Computer Science,

Indian Institute of Science Education and Research, Bhopal

Under the guidance of

Dr Parthiban Srinivasan

Acknowledgements

I would like to express my gratitude to Dr Parthiban Srinivasan, the course instructor for the course Data Science In Practice. His lectures helped me think practically about the applications of Data Science in solving domain specific problems. I enjoyed working on this project and learnt a lot about the theoretical underpinnings of Photometric classification. As a follow up of this project, I am deeply interested in exploring the use of Spiking Neural Networks and Physics Neural Networks in the same domain as I believe this would give me a chance in understanding the shortcomings of the current approaches used.

CONTENTS

Abstract

The Sloan Digital Sky Survey (SDSS) Data Release 17 (DR17) catalog contains photometric data for all objects viewed through a telescope and spectroscopic data for a small part of these. On the tagged photometric data that has been spectroscopically classified using labels, I trained ML classification models. The EDA aids in selecting the appropriate models for the dataset. One can choose a data model with the aid of all the knowledge we have obtained by conducting EDA. On fresh, unclassified data, I used the learned models with the highest accuracy on the training set. The three types of astronomical transients namely : stars, galaxies, and quasars have been predicted using a variety of machine learning methods. I have taken into account the classification techniques KNN, SVM, Random Forest and Decision Tree. Foraying into Neural Networks, I have also experimented with the Multilayer Perceptron Based Classifier. Before foraying into statistical analysis, a vigorous pictorial analysis was done through Violin Plots, KDE distributions, PDF function distributions and frequency polygons to name a

few. This model would be ideal to get a thorough understanding of the data visually, statistically, mathematically as well as analytically.

Introduction

One of the most fundamental classification schemes used in astronomy is that of galaxies, quasars, and stars. The brightest and farthest objects in the universe are quasars. The largest collections of stars are known as galaxies. About 500,000 observations (rows) of space make up the dataset. 18 feature columns are used to describe each observation, with the class column serving as the only dependent or goal variable. The result of the class column, i.e. whether it is a star, galaxy, or quasar, must also be determined.
Through a volume one hundred times bigger than what has been previously investigated, the SDSS provides a three-dimensional representation of the cosmos. The Sloan Digital Sky's fourth phase's final data release is called Data Release 17 (DR17). SDSS observations through January 2021 are included in DR17. The SDSS keeps track of the distances to 100,000 quasars, which are the farthest known objects, and provides us with a previously unheard-of suggestion about the distribution of stuff at the outer limits of the observable universe. The photographs produced by the SDSS will be significantly more sensitive and accurate than those from past surveys that relied on photographic plates since it is the first large-area survey to use electronic light detectors. The scientific community and the general public have access to the SDSS results electronically in the form of exact inventories of all things found as well as photographs of the objects found. By the time the survey was over, 15 terabytes of information had been generated overall.10,000 observations of space made by the SDSS make up the data. Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.
The dataset offers plenty of information about space to explore. Also, the class column is the perfect target for classification practices.

The classification models under consideration are :

1. k-Nearest Neighbours.
2. Decision Trees.
3. Random Forest.
4. Logistic Regression
5. Naive Bayesian classification
6. Support Vector Machines
7. Multi Layer Perceptron Classifier.

This study documents the importance of data and its interpretation in physics through familiarizing with state-of-the-art tools available in technology. The way we see data has changed over the centuries and yet the fundamentals of analysis remain more or less the same. Physics and science in general have always been a data driven field, from the early philosophers and astronomers looking up at the sky and tabulating the positions of planets and constellations.
Millions of petabytes of scientific data are being produced every single day and most of it is made available to the public through data sets and libraries.

.

History has witnessed how one can access terabytes of astronomical data sitting in their bedroom. Applying our own algorithms and regression on that data set can even lead to the discovery of new galaxy clusters and planets.

Math and statistics were the sole tools available to physicists for most of the history of human science, but the advent of transistor technology has resulted in such a quantum leap that programming now frequently outperforms Math in situations requiring big data sets. Languages like Python and R are the modern equivalent of the introduction of Calculus by Sir Isaac Newton. Having such versatile tools in one's arsenal will give further insight and ease of working with oceans of data that we can access now.

The aim of this study is to appreciate and comprehend some of the latest tools available in Machine Learning and Deep learning which can be used for analyzing physical data in Physics. For this we select some publicly available datasets and run our analysis on them and through the process, learn and appreciate the superiority of combining and utilizing technology in science.

The DataSet

The DataSets used in this study serve as the raw material on which the work is done. It is from publicly available sources and in fact the creators of the data encourage the scientific as well as the non-scientific community to access, explore and apply their intuitions using Machine learning and statistical tools.

Sloan Digital Sky Survey

What is the Sloan Digital Sky Survey?
The survey will map one-quarter of the entire sky in detail, determining the positions and absolute brightness of hundreds of millions of celestial objects. It will also measure the distances to more than a million galaxies and quasars.
The SDSS addresses fascinating, fundamental questions about the universe. With the survey, astronomers will be able to see the large-scale patterns of galaxies: sheets and voids through the whole universe. Scientists have many ideas about how the universe evolved, and different patterns of large-scale structure point to different theories. The Sloan Digital Sky Survey will tell us which theories are right - or whether we will have to come up with entirely new ideas.

Measuring Distance and Time: Redshift

The universe is expanding like a loaf of raisin bread rising in an oven. Pick any raisin, and imagine that it's our own Milky Way galaxy. If you place yourself on that raisin, then no matter how you look at the loaf, as the bread rises, all the other raisins move away from you. The farther away another raisin is from you, the faster it moves away. In the same way, all the other galaxies are moving away from ours as the universe expands. And because the universe is uniformly expanding, the farther a galaxy is from Earth, the faster it is receding from us. The light coming to us from these distant objects is shifted toward the red end of the electromagnetic spectrum, in much the same way the sound of a train whistle changes as a train leaves or approaches a station. The faster a distant object is moving, the more it is redshifted. Astronomers measure the amount of redshift in the spectrum of a galaxy to figure out how far away it is from us.

By measuring the redshifts of a million galaxies, the Sloan Digital Sky Survey will provide a three-dimensional picture of our local neighborhood of the universe.

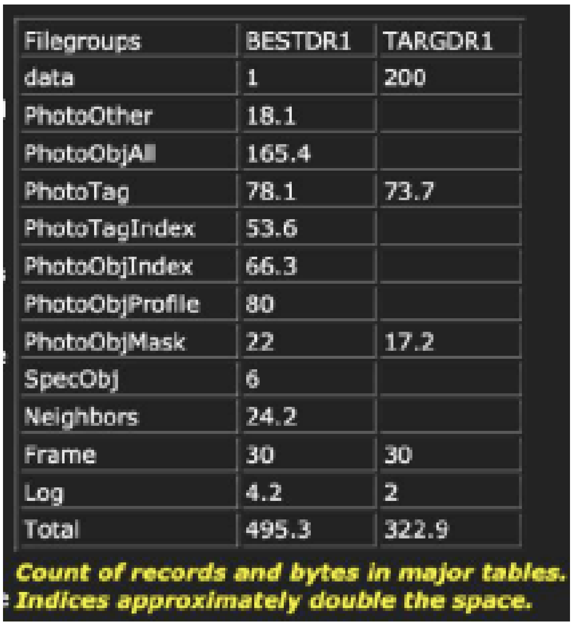
The Databases

The processed data are stored in databases. The logical database design consists of photographic and spectrographic objects. They are organized into a pair of snowflake schemas. Sub setting views and many indices give convenient access to the conventional subsets (such as stars and galaxies). Procedures and indices are defined to make spatial lookups convenient and fast.

Database Physical Design

SkyServer initially took a simple approach to database design – and since that worked, we stopped there. The design counts on the SQL storage engine and query optimizer to make all the intelligent decisions about data layout and data access.

The total amount of data in the two databases is 818 GB, and the total number of rows exceeds 3.4 billion. The data tables are all created in several filegroups. The database files are spread across a single RAID0 volume. Each filegroup contains several database files that are limited to about 50Gb each. The log files and temporary database are also spread across these disks. SQL Server stripes the tables across all these files and hence across all these disks. It detects the sequential access, creates the parallel prefetch threads, and uses multiple processors to analyze the data as quickly as the disks can produce it. When reading or writing, this automatically gives the sum of the disk bandwidths (over 400 MBps peak, 180MBps typical) without any special user programming.



Filegroups	BESTDR1	TARGDR1
data	1	200
PhotoOther	18.1	
PhotoObjAll	165.4	
PhotoTag	78.1	73.7
PhotoTagIndex	53.6	
PhotoObjIndex	66.3	
PhotoObjProfile	80	
PhotoObjMask	22	17.2
SpecObj	6	
Neighbors	24.2	
Frame	30	30
Log	4.2	2
Total	495.3	322.9

Count of records and bytes in major tables. Indices approximately double the space.

Beyond this file group striping, SkyServer uses all the SQL Server default values. There is no special tuning. This is the hallmark of SQL Server – the system aims to have "no knobs" so that the out-of-the box performance is quite good. The SkyServer is a testimonial to that goal.

Metadata about DR 17 used in this project

The above section gave a broad overview about the objectives and methods used in Sloan Digital Sky Survey. We use only a small subset of the main dataset and our algorithms and visualizations will be done in that subset. The data released by the SDSS is under public domain. It's taken from the current data release DR17.

The data consists of 10,000 observations of space taken by the SDSS. Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.

Inspiration: The dataset offers plenty of information about space to explore. Also, the class column is the perfect target for classification practices.

Feature Description

The table results from a query which joins two tables (actually views): "PhotoObj" which contains photometric data and "SpecObj" which contains spectral data.

The feature descriptions are as below:

View "PhotoObj"

- obj_ID = Object Identifier
- alpha = J2000 Right Ascension (r-band)
- delta = J2000 Declination (r-band)

Right ascension (abbreviated RA) is the angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth in question. When paired with declination (abbreviated dec), these astronomical coordinates specify the direction of a point on the celestial sphere (traditionally called in English the skies or the sky) in the equatorial coordinate system.

Source: https://en.wikipedia.org/wiki/Right_ascension

- u = better of DeV/Exp magnitude fit
- g = better of DeV/Exp magnitude fit
- r = better of DeV/Exp magnitude fit
- i = better of DeV/Exp magnitude fit
- z = better of DeV/Exp magnitude fit

The Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope.

Further education: https://www.astro.umd.edu/~ssm/ASTR620/mags.html

- run_ID = Run Number
- rereun_ID = Rerun Number
- camcol_ID = Camera column
- field_ID = Field number

Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by:

- run number, which identifies the specific scan,
- the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and

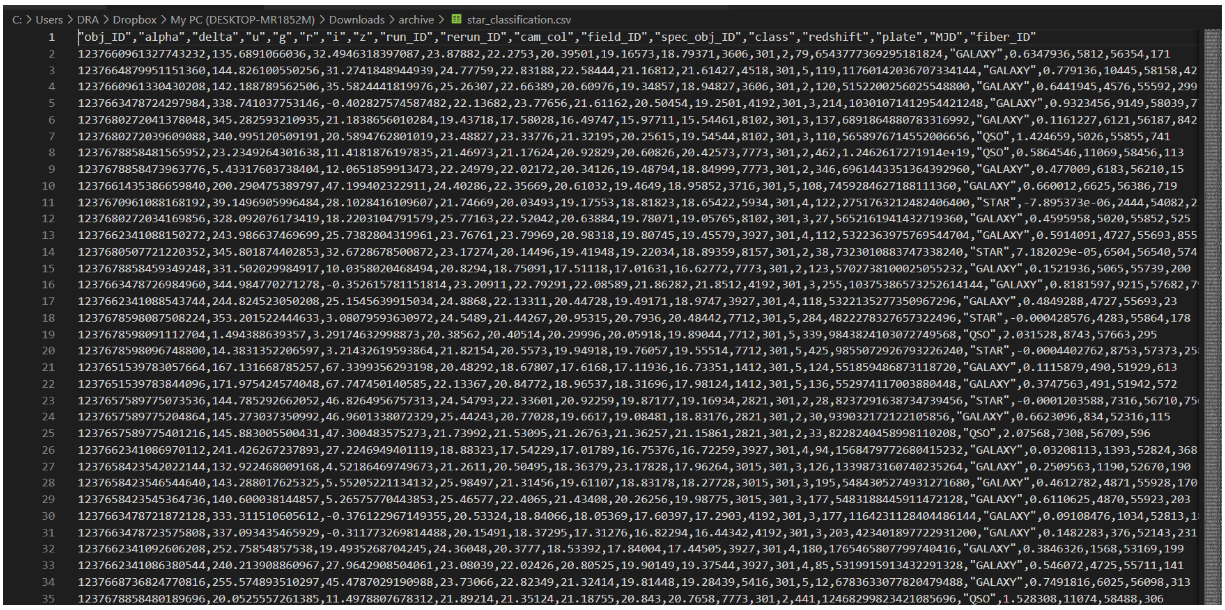
- the field number. The field number typically starts at 11 (after an initial rampup time), and can be as large as 800 for particularly long runs.
- An additional number, rerun, specifies how the image was processed. View "SpecObj"
- specobj_ID = Object Identifier
- class = object class (galaxy, star or quasar object)

The class identifies an object to be either a galaxy, star or quasar. This will be the response variable which we will be trying to predict.

- redshift = Final Redshift
- plate = plate number
- MJD = MJD of observation
- fiber_ID = fiber ID

In physics, redshift happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.

Each spectroscopic exposure employs a large, thin, circular metal plate that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane. These fibers then feed into the spectrographs. Each plate has a unique serial number, which is called plate in views such as SpecObj in the CAS.
Modified Julian Date, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.
The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID.
The DataSet for our use can be downloaded as a CSV file which can be opened using excel. CSV stands for 'comma separated values' and it's a format for storing tabular data. The DataSet will have 100,000 entries ad 17 feature columns and 1 class column. This can be seen while opening the file (I have opened it in VS Code)


C: > Users > DRA > Dropbox > My PC (DESKTOP-MR1852M) > Downloads > archive > star_classification.csv

```
"obj_ID","alpha","delta","u","g","r","i","z","run_ID","rerun_ID","cam_col","field_ID","spec_obj_ID","class","redshift","plate","MJD","fiber_ID"
1237660961327743232,135.6891066036,32.4946318397087,23.87882,22.2753,20.39501,19.16573,18.79371,3606,301,2,79,6543777369295181824,"GALAXY",0.6347936,5812,56354,171
1237664879951151360,144.826100550256,31.2741848944939,24.77759,22.83188,22.58444,21.16812,21.61427,4518,301,5,119,11760142036707334144,"GALAXY",0.779136,10445,58158,42
1237660961330430208,142.188789562506,35.5824441819976,25.26307,22.66389,20.60976,19.34857,18.94827,3606,301,2,120,5152200256025548800,"GALAXY",0.6441945,4576,55592,299
1237663478724297984,338.741037753146,-0.402827574587482,22.13682,23.77656,21.61162,20.50454,19.2501,4192,301,3,214,10301071412954421248,"GALAXY",0.9323456,9149,58039,7
1237680272041378048,345.282593210935,21.1838656010284,19.43718,17.58028,16.49747,15.97711,15.54461,8102,301,3,137,6891864880783316992,"GALAXY",0.1161227,6121,56187,842
1237680272039609088,340.995120509191,20.5894762801019,23.48827,23.33776,21.32195,20.25615,19.54544,8102,301,3,110,5658976714552006656,"QSO",1.424659,5026,55855,741
1237678858481565952,23.2349264301638,11.4181876197835,21.46973,21.17624,20.92829,20.60826,20.42573,7773,301,2,462,1.2462617271914e+19,"QSO",0.5864546,11069,58456,113
1237678858473963776,5.43317603738404,12.0651859913473,22.24979,22.02172,20.34126,19.48794,18.84999,7773,301,2,346,6961443351364392960,"GALAXY",0.477009,6183,56210,15
1237661435386659840,200.290475389797,47.199402322911,24.40286,22.35669,20.61032,19.4649,18.95852,3716,301,5,108,7459284627188111360,"GALAXY",0.660012,6625,56386,719
1237670961088168192,39.1496905996484,28.1028416109607,21.74669,20.03493,19.17553,18.81823,18.65422,5934,301,4,122,2751763212482406400,"STAR",-7.895373e-06,2444,54082,2
1237680272034169856,328.092076173419,18.2203104791579,25.77163,22.52042,20.63884,19.78071,19.05765,8102,301,3,27,5652161941432719360,"GALAXY",0.4595958,5020,55852,525
1237662341088150272,243.986637469699,25.7382804319961,23.76761,23.79969,20.98318,19.80745,19.45579,3927,301,4,112,5322363975769544704,"GALAXY",0.5914091,4727,55693,855
1237680507721220352,345.801874402853,32.6728678500872,23.17274,20.14496,19.41948,19.22034,18.89359,8157,301,2,38,7323010883747338240,"STAR",7.182029e-05,6504,56540,574
1237678858459349248,331.502029984917,10.0358020468494,20.8294,18.75091,17.51118,17.01631,16.62772,7773,301,2,123,5702738100025055232,"GALAXY",0.1521936,5065,55739,200
1237663478726984960,344.984770271278,-0.352615781151814,23.20911,22.79291,22.08589,21.86282,21.8512,4192,301,3,255,10375386573252614144,"GALAXY",0.8181597,9215,57682,7
1237662341088543744,244.824523050208,25.1545639915034,24.8868,22.13311,20.44728,19.49171,18.9747,3927,301,4,118,5322135277350967296,"GALAXY",0.4849288,4727,55693,23
1237678598087508224,353.201522444633,3.08079593630972,24.5489,21.44267,20.95315,20.7936,20.48442,7712,301,5,284,4822278327657322496,"STAR",-0.000428576,4283,55864,178
1237678598091112704,1.494388639357,3.29174632998873,20.38562,20.40514,20.29996,20.05918,19.89044,7712,301,5,339,9843824103072749568,"QSO",2.031528,8743,57663,295
1237678598096748800,14.3831352206597,3.21432619593864,21.82154,20.5573,19.94918,19.76057,19.55514,7712,301,5,425,9855072926793226240,"STAR",-0.0004402762,8753,57373,25
1237651539783057664,167.131668785257,67.3399356293198,20.48292,18.67807,17.6168,17.11936,16.73351,1412,301,5,124,5518594868731187 20,"GALAXY",0.1115879,490,51929,613
1237651539783844096,171.975424574048,67.747450140585,22.13367,20.84772,18.96537,18.31696,17.98124,1412,301,5,136,5529741170038804 48,"GALAXY",0.3747563,491,51942,572
1237657589775073536,144.785292662052,46.8264956757313,24.54793,22.33601,20.92259,19.87177,19.16934,2821,301,2,28,8237291638734739456,"STAR",-0.0001203588,7316,56710,75
1237657589775204864,145.273037350992,46.9601338072329,25.44243,20.77028,19.6617,19.08481,18.83176,2821,301,2,30,9390321721221058 56,"GALAXY",0.6623096,834,52316,115
1237657589775401216,145.883005500431,47.300483575273,21.73992,21.53095,21.26763,21.36257,21.15861,2821,301,2,33,8228240458998110208,"QSO",2.07568,7308,56709,596
1237662341086970112,241.426267237893,27.2246949401119,18.88323,17.54229,17.01789,16.75376,16.72259,3927,301,4,94,1568479772680415232,"GALAXY",0.03208113,1393,52824,368
1237658423542022144,132.922468009168,4.52186469749673,21.2611,20.50495,18.36379,23.17828,17.96264,3015,301,3,126,1339873160740235264,"GALAXY",0.2509563,1190,52670,190
1237658423546544640,143.288017625325,5.55205221134132,25.98497,21.31456,19.61107,18.83178,18.27728,3015,301,3,195,5484305274931271680,"GALAXY",0.4612782,4871,55928,170
1237658423545364736,140.600038144857,5.26575770443853,25.46577,22.4065,21.43408,20.26256,19.98775,3015,301,3,177,5483188445911472128,"GALAXY",0.6110625,4870,55923,203
1237663478721872128,333.311510605612,-0.376122967149355,20.53324,18.84066,18.05369,17.60397,17.2903,4192,301,3,177,1164231128404486144,"GALAXY",0.09108476,1034,52813,1
1237663478723575808,337.093435465929,-0.311773269814488,20.15491,18.37295,17.31276,16.82294,16.44342,4192,301,3,203,423401897722931200,"GALAXY",0.1482283,376,52143,231
1237662341092606208,252.75854857538,19.4935268704245,24.36048,20.3777,18.53392,17.84004,17.44505,3927,301,4,180,1765465807799740416,"GALAXY",0.3846326,1568,53169,199
1237662341086380544,240.213908860967,27.9642908504061,23.08039,22.02426,20.80525,19.90149,19.37544,3927,301,4,85,5319915913432291328,"GALAXY",0.546072,4725,55711,141
1237668736824770816,255.574893510297,45.4787029190988,23.73066,22.82349,21.32414,19.81448,19.28439,5416,301,5,12,6783633077820479488,"GALAXY",0.7491816,6025,56098,313
1237678858480189696,20.0525557261385,11.4978807678312,21.89214,21.35124,21.18755,20.843,20.7658,7773,301,2,441,1246829982342108 5696,"QSO",1.528308,11074,58488,306
```

Structural flow deployed in the project

1. ENVIRONMENT AND PARAMETERS CONFIGURATION

1.1 Importing required libraries

2. DATA PREPARATION

2.1 Loading the Data

2.2 Dropping Features that are not significant

3. HIGH LEVEL STATISTICS

3.1 Describing the Data

3.2 Finding Duplicates

3.3 Finding Unique Values

3.4 Finding Null Values

4. UNIVARIATE ANALYSIS

4.1 BoxPlot

4.2 Frequency Polygon

4.3 Histplot

4.4 Probability Density Function (PDF)

4.5 Violin Plot

4.6 KDE Plot

5. MULTIVARIATE ANALYSIS

5.1 LinePlot

5.2 Correlation between variables

5.3 3D Scatter Plot

5.4 Pairplots

5.5 Pandas Profiling

5.6 Equatorial Coordinates Plot

6. ML Algorithms

6.1 Feature Engineering

6.2 Logistic Regression

6.3 KNN

6.4 Naïve Bayes

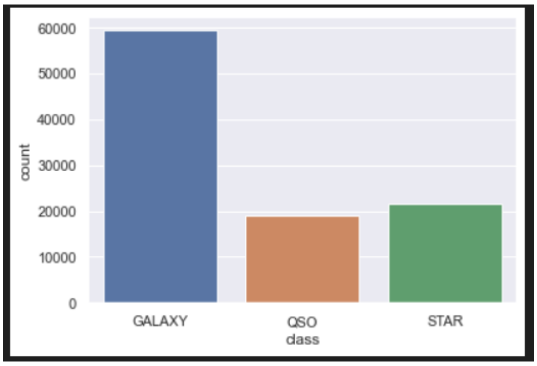
6.5 Random Forest

6.7 SVM

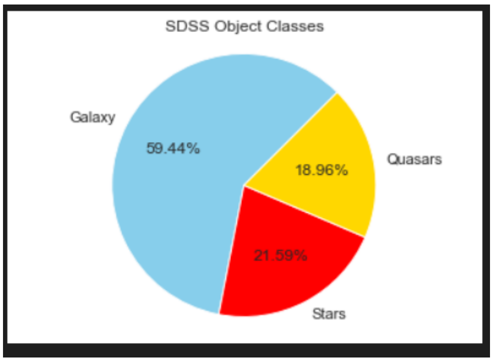
6.8 Multi-Layer Perceptron Classifier (Neural Network)

6.9 Decision Tree

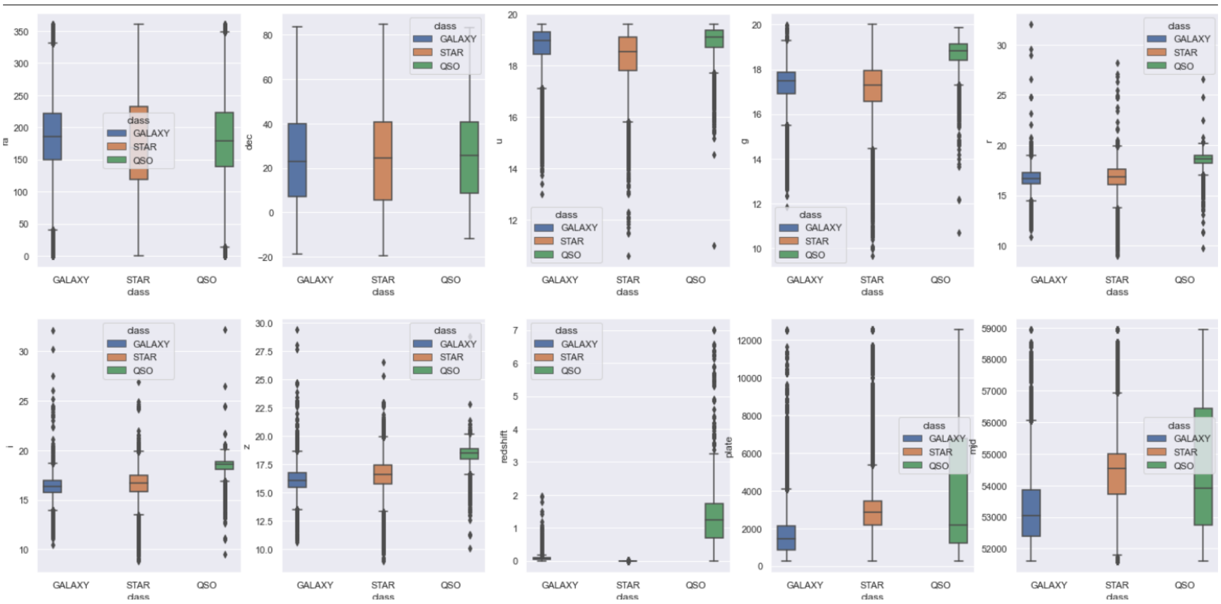
Following are some of the plots obtained from EDA



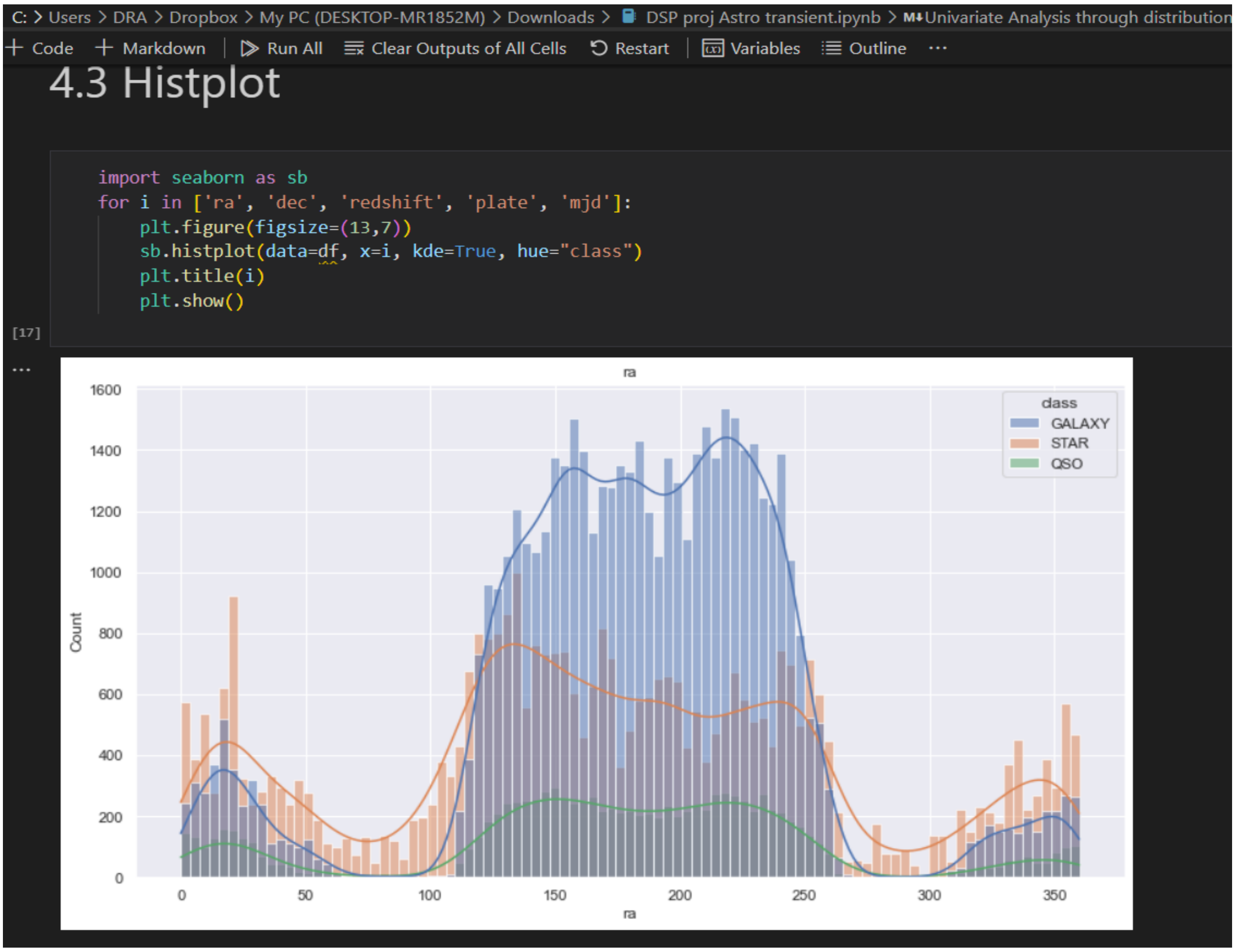
Bar Plot



Pie Chart



Box Plot

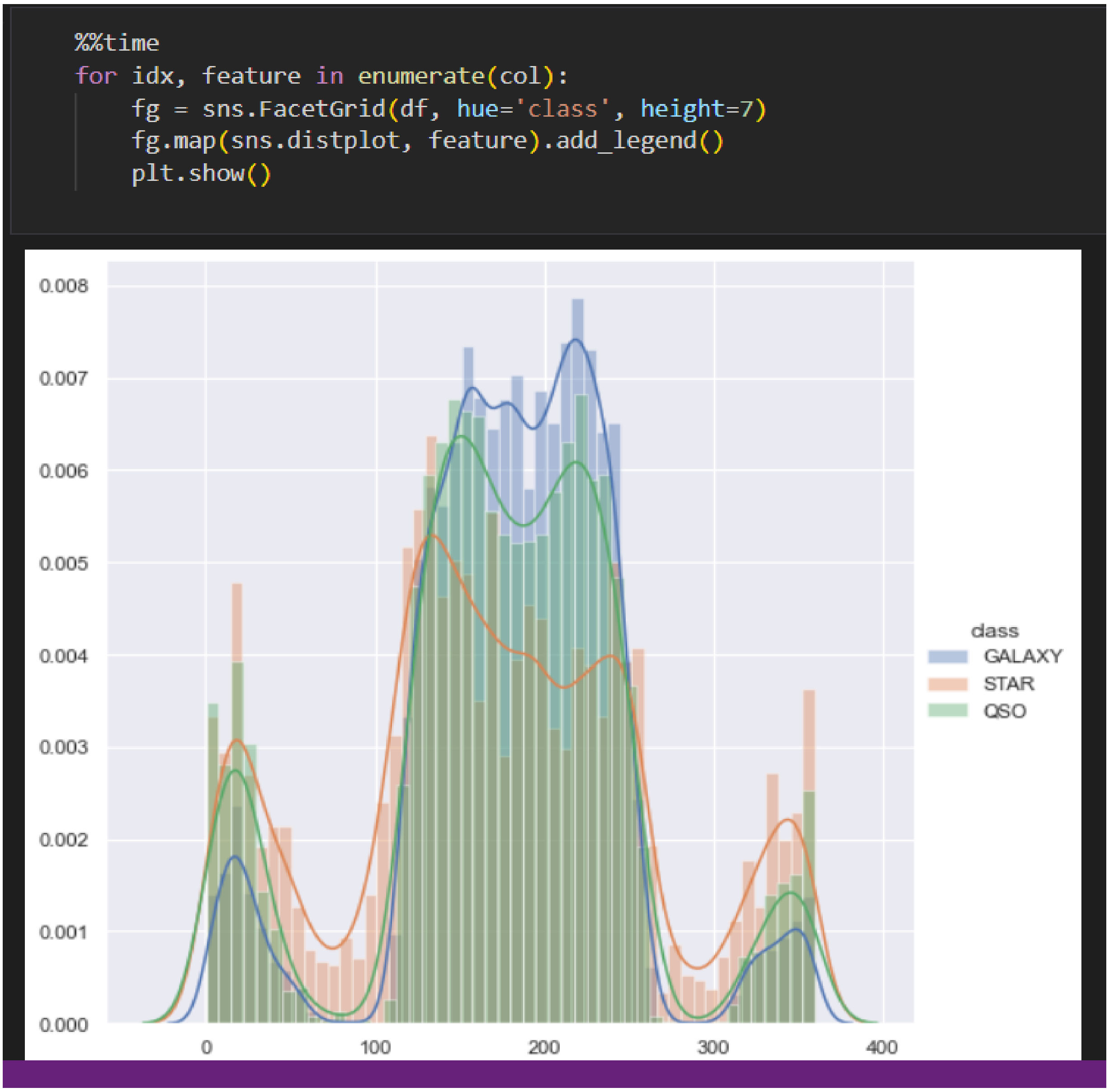

C: > Users > DRA > Dropbox > My PC (DESKTOP-MR1852M) > Downloads > DSP proj Astro transient.ipynb > M↓Univariate Analysis through distribution
+ Code
+ Markdown
▷ Run All
Clear Outputs of All Cells
Restart
Variables
Outline
···

4.3 Histplot

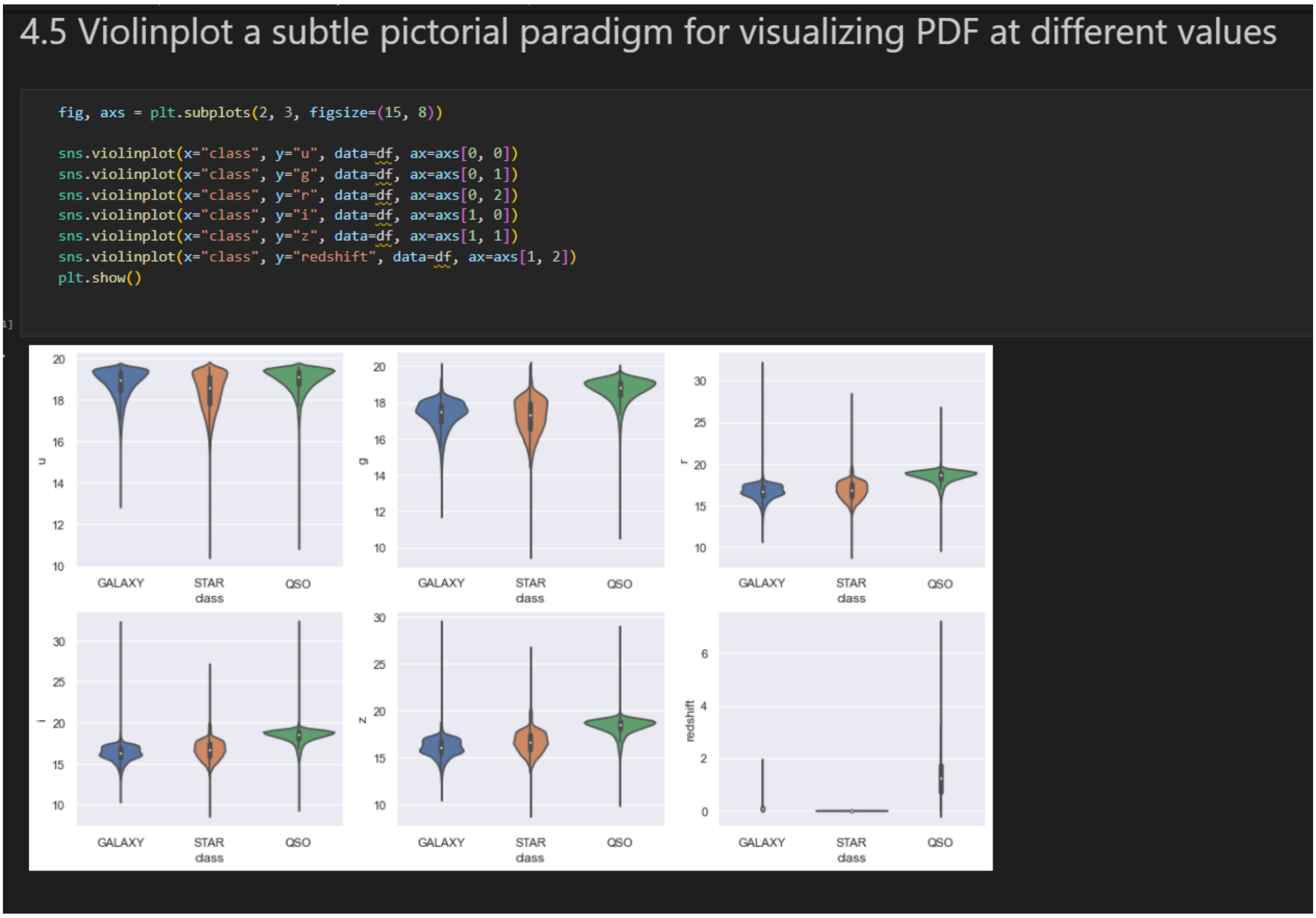
```
import seaborn as sb
for i in ['ra', 'dec', 'redshift', 'plate', 'mjd']:
    plt.figure(figsize=(13,7))
    sb.histplot(data=df, x=i, kde=True, hue="class")
    plt.title(i)
    plt.show()
```

[17]

ra
class
GALAXY
STAR
QSO
Count
1600
1400
1200
1000
800
600
400
200
0
0
50
100
150
200
250
300
350
ra

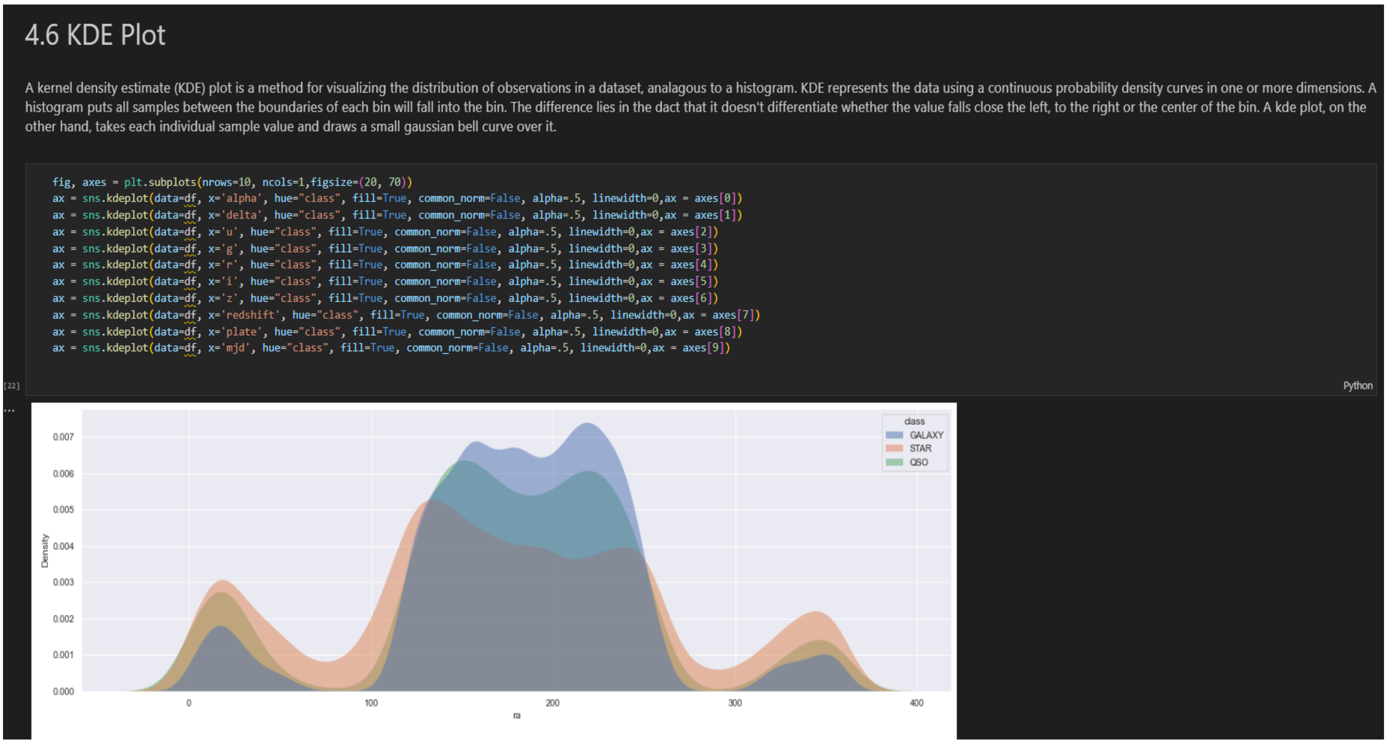


```
%%time
for idx, feature in enumerate(col):
    fg = sns.FacetGrid(df, hue='class', height=7)
    fg.map(sns.distplot, feature).add_legend()
    plt.show()
```



4.5 Violinplot a subtle pictorial paradigm for visualizing PDF at different values

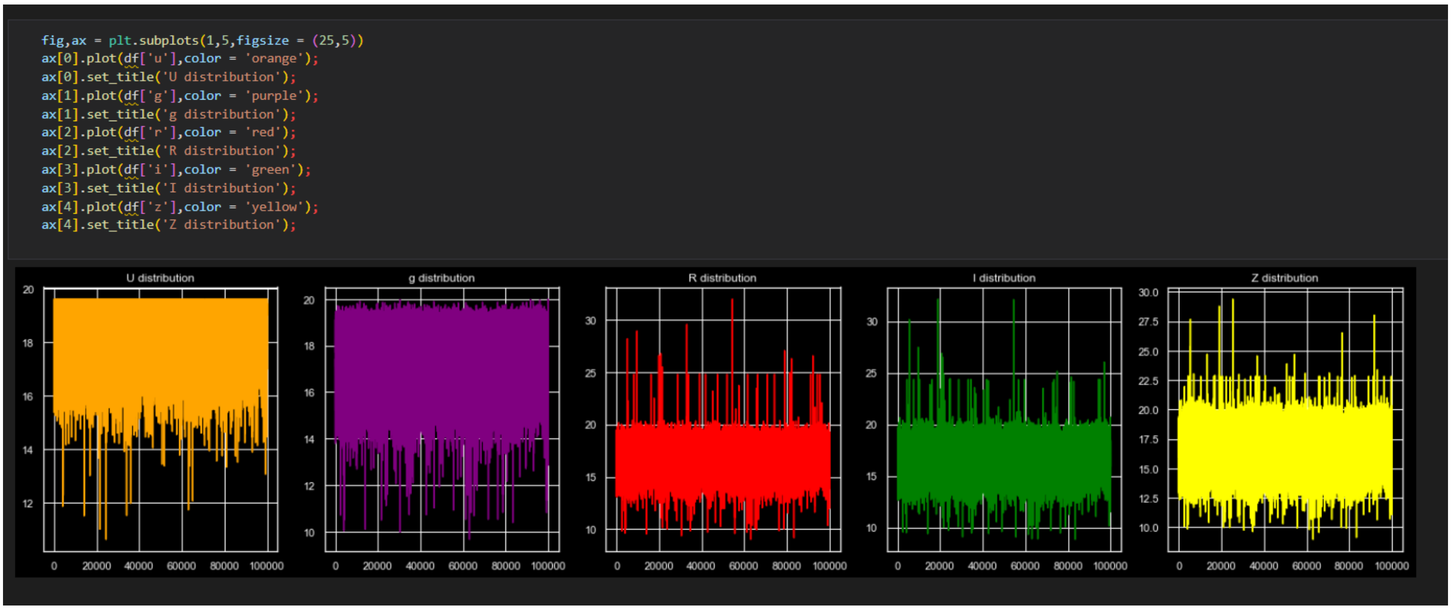
```
fig, axs = plt.subplots(2, 3, figsize=(15, 8))

sns.violinplot(x="class", y="u", data=df, ax=axs[0, 0])
sns.violinplot(x="class", y="g", data=df, ax=axs[0, 1])
sns.violinplot(x="class", y="r", data=df, ax=axs[0, 2])
sns.violinplot(x="class", y="i", data=df, ax=axs[1, 0])
sns.violinplot(x="class", y="z", data=df, ax=axs[1, 1])
sns.violinplot(x="class", y="redshift", data=df, ax=axs[1, 2])
plt.show()
```



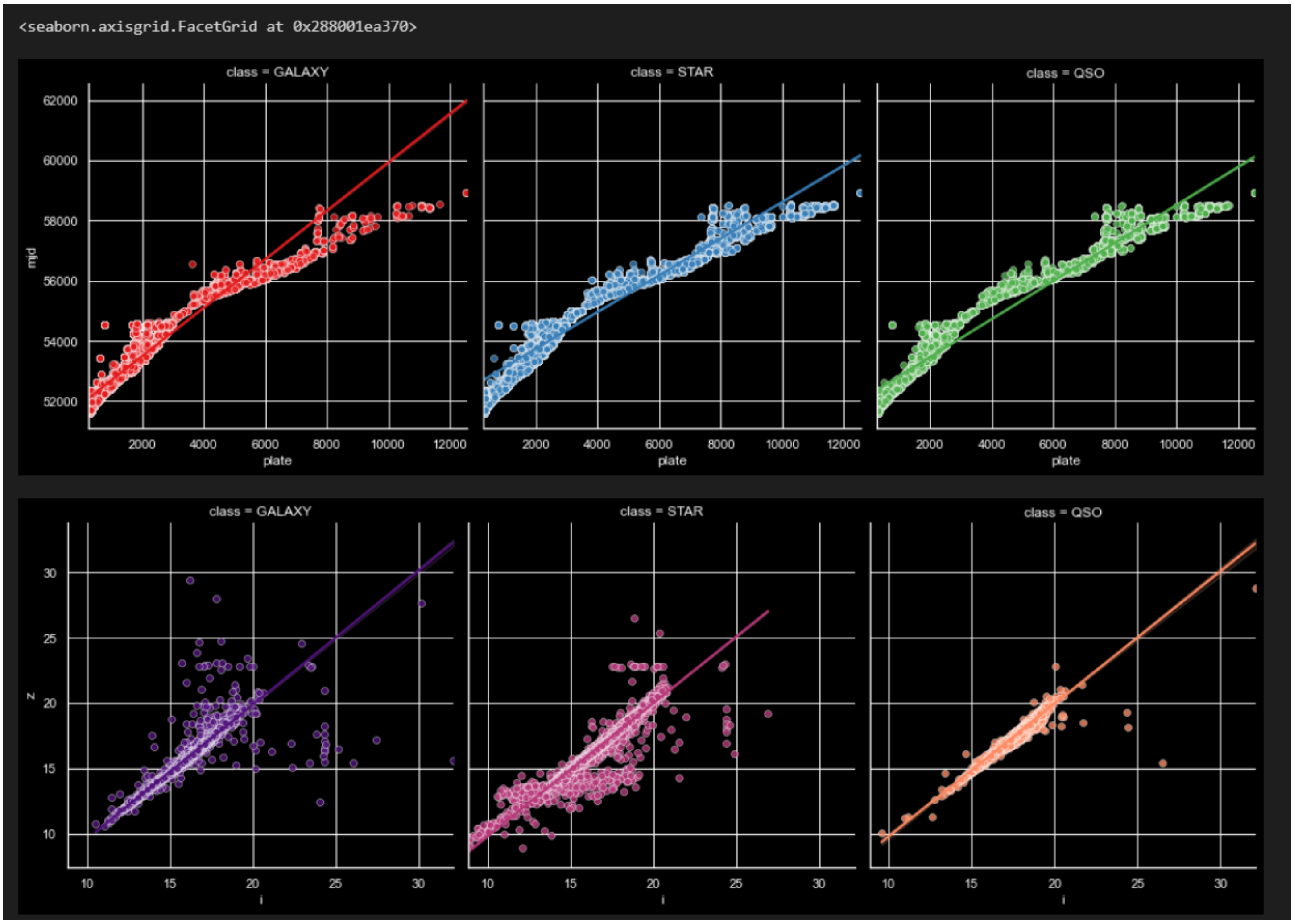
4.6 KDE Plot

A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset, analagous to a histogram. KDE represents the data using a continuous probability density curves in one or more dimensions. A histogram puts all samples between the boundaries of each bin will fall into the bin. The difference lies in the dact that it doesn't differentiate whether the value falls close the left, to the right or the center of the bin. A kde plot, on the other hand, takes each individual sample value and draws a small gaussian bell curve over it.

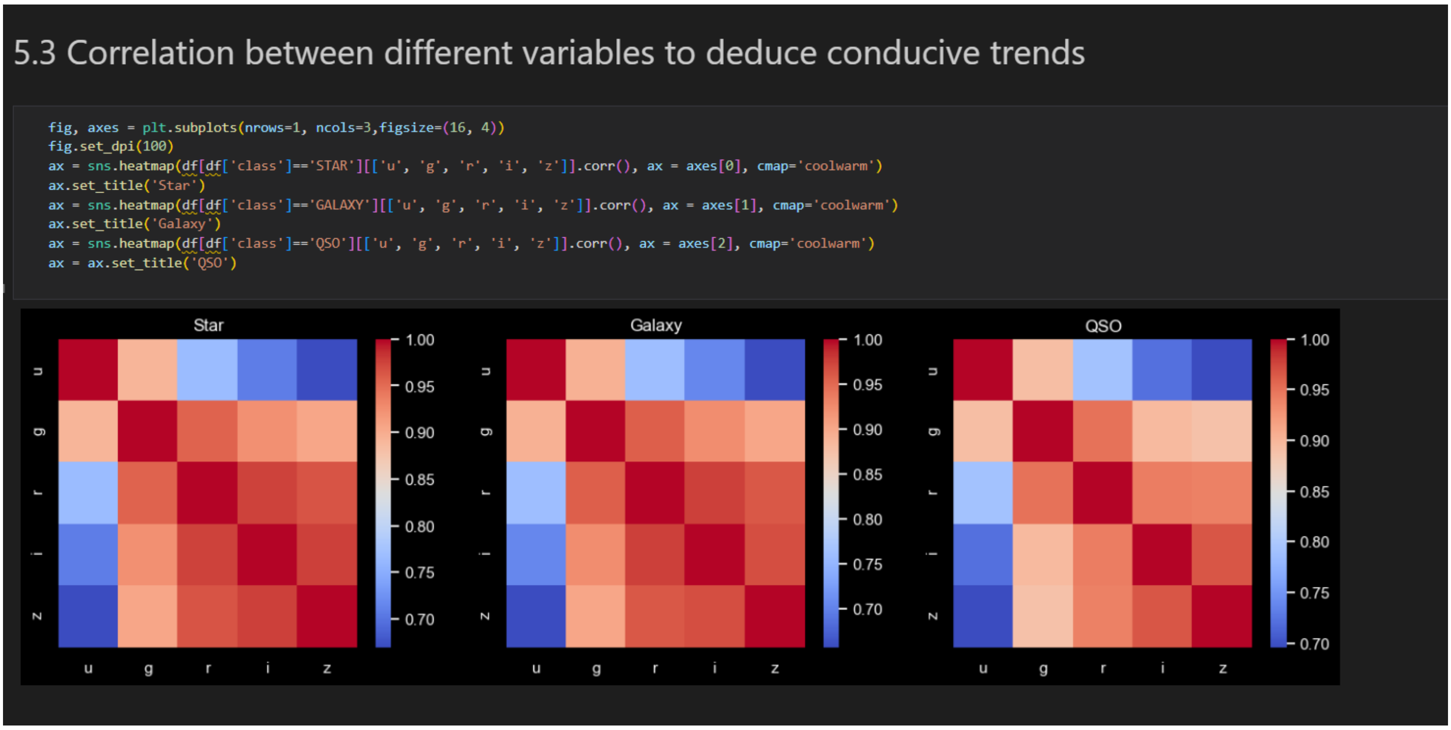
```
fig, axes = plt.subplots(nrows=10, ncols=1,figsize=(20, 70))
ax = sns.kdeplot(data=df, x='alpha', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[0])
ax = sns.kdeplot(data=df, x='delta', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[1])
ax = sns.kdeplot(data=df, x='u', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[2])
ax = sns.kdeplot(data=df, x='g', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[3])
ax = sns.kdeplot(data=df, x='r', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[4])
ax = sns.kdeplot(data=df, x='i', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[5])
ax = sns.kdeplot(data=df, x='z', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[6])
ax = sns.kdeplot(data=df, x='redshift', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[7])
ax = sns.kdeplot(data=df, x='plate', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[8])
ax = sns.kdeplot(data=df, x='mjd', hue="class", fill=True, common_norm=False, alpha=.5, linewidth=0,ax = axes[9])
```



```
fig,ax = plt.subplots(1,5,figsize = (25,5))
ax[0].plot(df['u'],color = 'orange');
ax[0].set_title('U distribution');
ax[1].plot(df['g'],color = 'purple');
ax[1].set_title('g distribution');
ax[2].plot(df['r'],color = 'red');
ax[2].set_title('R distribution');
ax[3].plot(df['i'],color = 'green');
ax[3].set_title('I distribution');
ax[4].plot(df['z'],color = 'yellow');
ax[4].set_title('Z distribution');
```

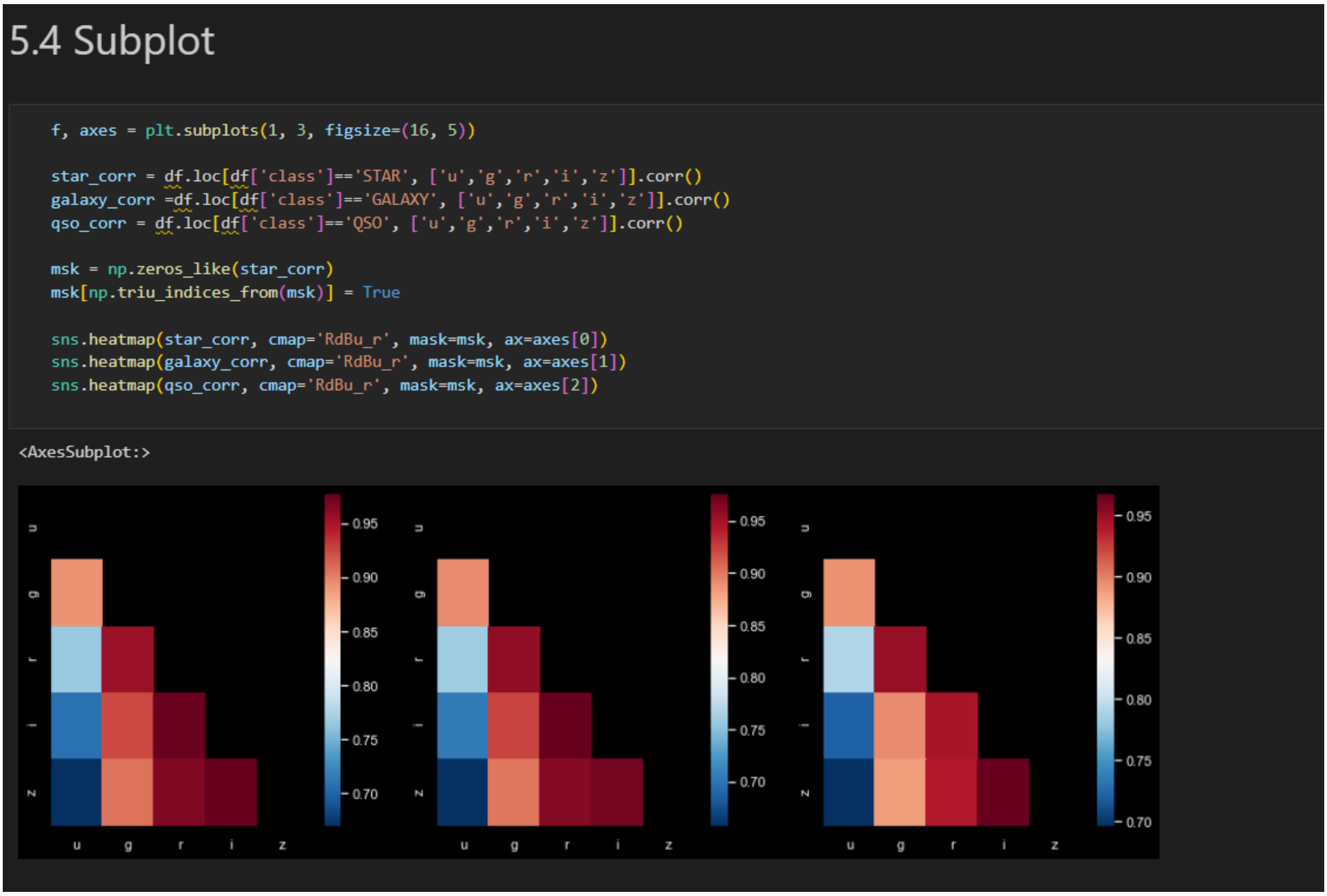


```
<seaborn.axisgrid.FacetGrid at 0x288001ea370>
```



5.3 Correlation between different variables to deduce conducive trends

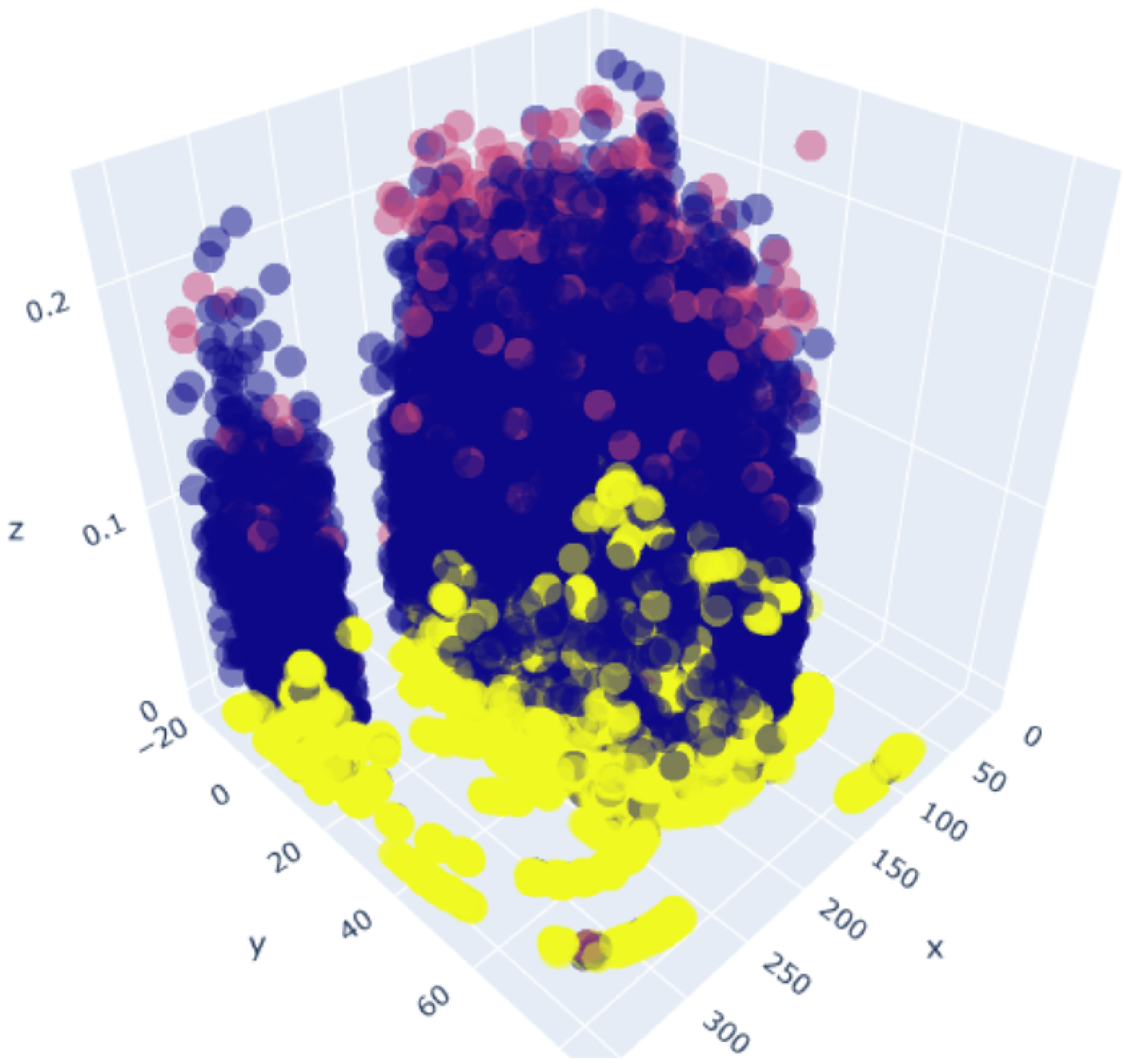
```
fig, axes = plt.subplots(nrows=1, ncols=3,figsize=(16, 4))
fig.set_dpi(100)
ax = sns.heatmap(df[df['class']=='STAR'][['u', 'g', 'r', 'i', 'z']].corr(), ax = axes[0], cmap='coolwarm')
ax.set_title('Star')
ax = sns.heatmap(df[df['class']=='GALAXY'][['u', 'g', 'r', 'i', 'z']].corr(), ax = axes[1], cmap='coolwarm')
ax.set_title('Galaxy')
ax = sns.heatmap(df[df['class']=='QSO'][['u', 'g', 'r', 'i', 'z']].corr(), ax = axes[2], cmap='coolwarm')
ax = ax.set_title('QSO')
```

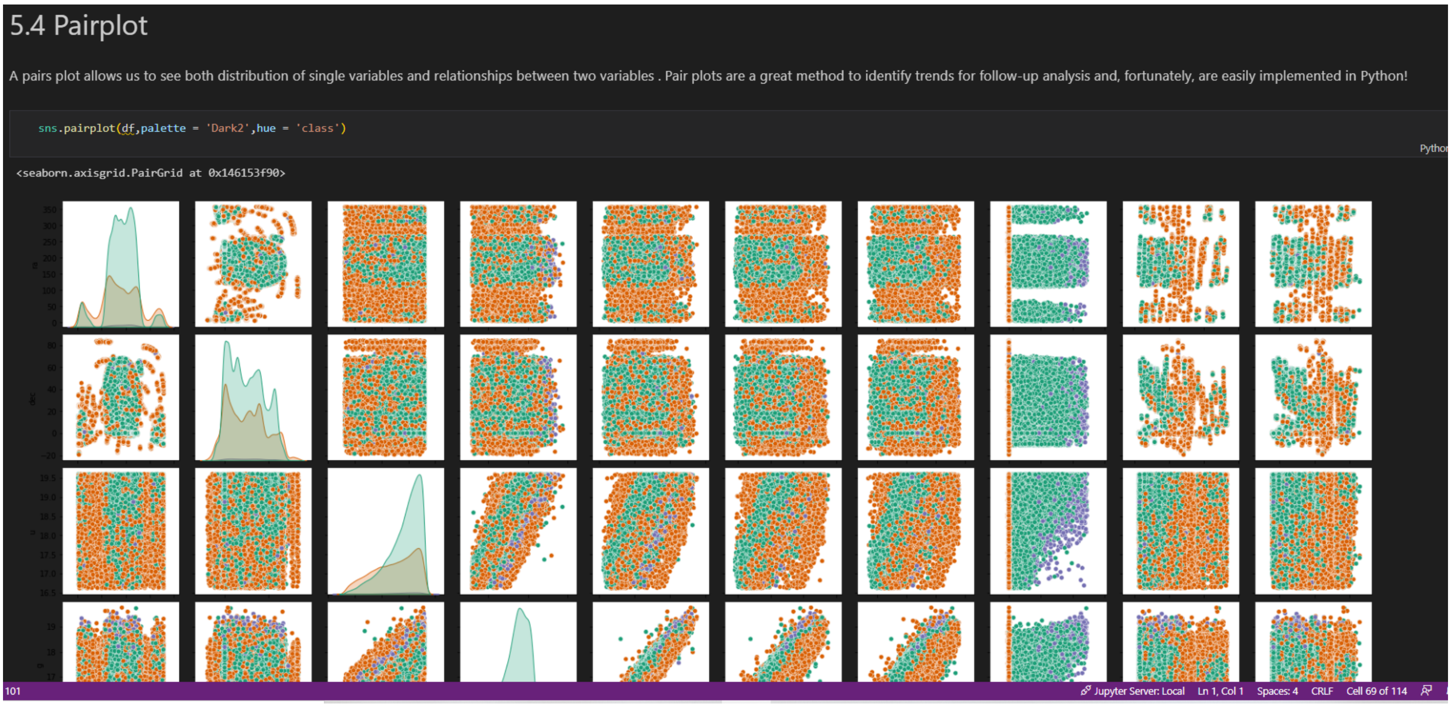


5.4 Subplot

```
f, axes = plt.subplots(1, 3, figsize=(16, 5))

star_corr = df.loc[df['class']=='STAR', ['u','g','r','i','z']].corr()
galaxy_corr =df.loc[df['class']=='GALAXY', ['u','g','r','i','z']].corr()
qso_corr = df.loc[df['class']=='QSO', ['u','g','r','i','z']].corr()

msk = np.zeros_like(star_corr)
msk[np.triu_indices_from(msk)] = True

sns.heatmap(star_corr, cmap='RdBu_r', mask=msk, ax=axes[0])
sns.heatmap(galaxy_corr, cmap='RdBu_r', mask=msk, ax=axes[1])
sns.heatmap(qso_corr, cmap='RdBu_r', mask=msk, ax=axes[2])
```

```
<AxesSubplot:>
```



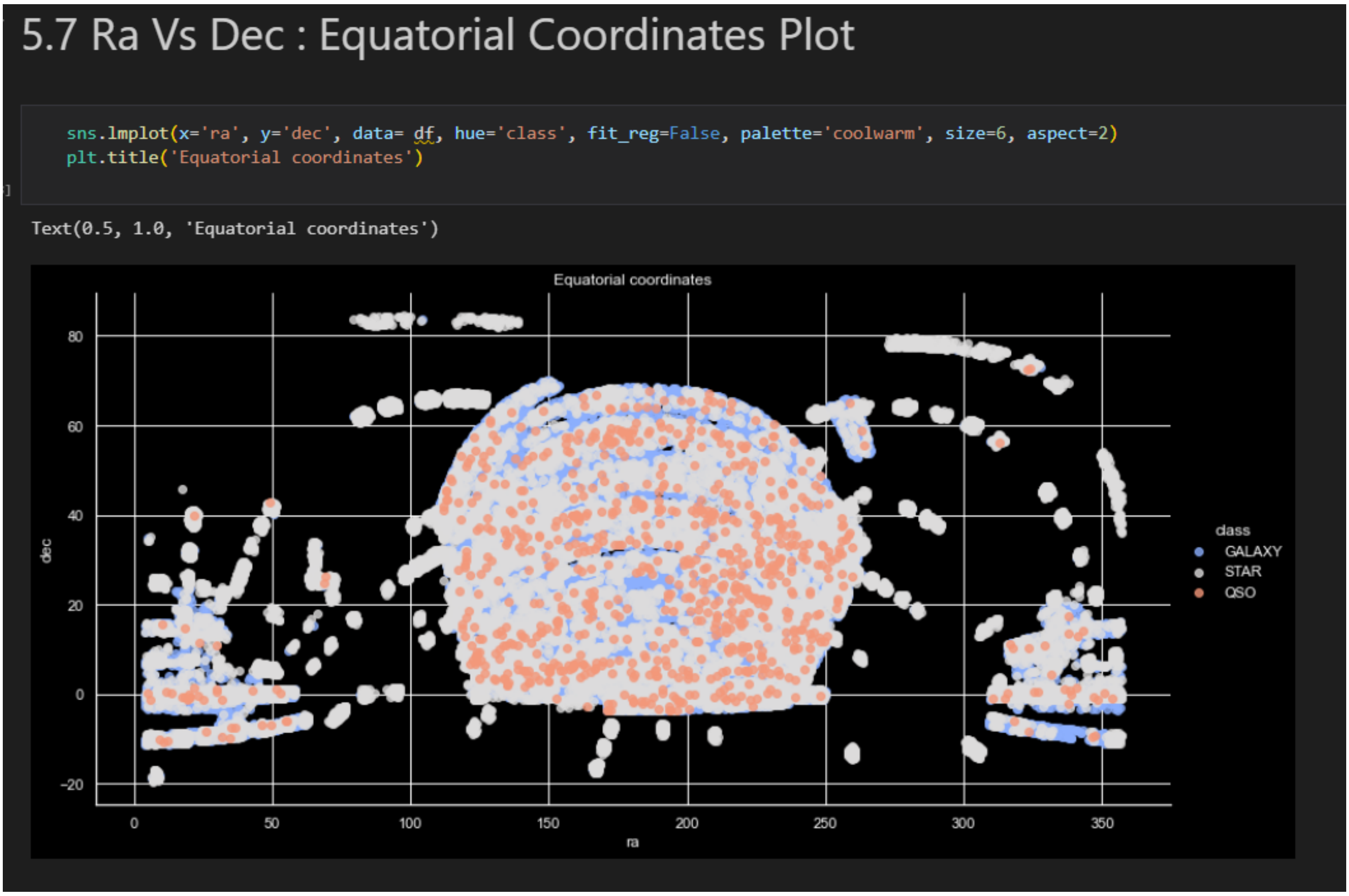


5.4 Pairplot

A pairs plot allows us to see both distribution of single variables and relationships between two variables . Pair plots are a great method to identify trends for follow-up analysis and, fortunately, are easily implemented in Python!

```
sns.pairplot(df,palette = 'Dark2',hue = 'class')
```

```
<seaborn.axisgrid.PairGrid at 0x146153f90>
```



5.7 Ra Vs Dec : Equatorial Coordinates Plot

```
sns.lmplot(x='ra', y='dec', data= df, hue='class', fit_reg=False, palette='coolwarm', size=6, aspect=2)
plt.title('Equatorial coordinates')
```

```
Text(0.5, 1.0, 'Equatorial coordinates')
```

Analysis of Results and Discussion

The following can be concluded from the Exploratory Data Analysis:

1. Stars have the lowest average redshift, followed by Galaxies and then Quasars.
2. Hotter objects emit more of every wavelength. In tandem with the theory, u,g,r,i,z correlation looks in accordance with expected physical behavior
3. All wavelength radiations are strongly correlated except ‘u’ owing to lower correlations that can be observed from the frequency polygon and kde distribution.
4. From the violin plot it can be said that, value of obj_ID and alpha don't contribute to classifying Star or Galaxy implying that classification doesn’t depend on the right ascension angle
5. In the 'redshift' feature, if the value is negative (blueshift), the observation is more likely to be a Star. If the value is positive (redshift), the observation is more likely to be Galaxy.

The following can be concluded from the ML and DL statistical analysis section

1. Random Forest scores the best accuracy and the minimum deviation, hence this could be considered an apt model for the photometric task at hand
2. Galaxies are easier to separate than stars and quasars from the inferences of the General Theory of relativity.
3. Although quasars and stars might show resemblance quantitatively and graphically from the PDFs, they have some distinctive statistical distributions innate to the geometries that help filters classify them.

S.no	Model name	Accuracy (in %)
1	Random Forest	98.96
2	Decision Tree	98.73
3	SVM	97.56
4	KNN Neighbors	87.41
5	Logistic Regression	89.94
6	Naive Bayes	97.15
7	Multi layer Perceptron Classifier	98.60

Conclusion

My insights on science and the true work that scientists do have culminated in this effort. In some ways, this study has been successful in giving an overview of the methods and tools used by researchers to tackle challenges in the actual world. This is a feeble attempt to depict a sample of the actual scientific investigation taking place throughout the globe even as this is being written.

In technical terms, to summarize

- The application of EDA on Astronomy DataSet from SDSS has given insights into how photometric parameters (FEATURES) contribute to class labels(STAR< GALAXY, QUASAR)

- EDA with Statistical summaries as well as extensive visualizations provided the guidelines to transform the data as input for ML models.
- A comparison of the accuracy and performance of the different ML models were done.

References

1. The data released by the SDSS is under public domain. It’s taken from the current data release RD14. More information about the license:
http://www.sdss.org/science/image-gallery/
2. It was acquired by querying the CasJobs database which contains all the data published by the SDSS. The exact query can be found at:
http://skyserver.sdss.org/CasJobs/ (Free account is required)
3. There are also other ways to get data from the SDSS catalogue. They can be found under:
http://www.sdss.org/dr14/
4. Scikit-Learn Implementation of SVM:
https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
5. https://medium.com/@suvigya2001/the-gaussian-rbf-kernel-in-non-linear svm-2fb1c822aae0
6. https://engineering.papercup.com/posts/kernel-methods/
7. https://machinelearningmastery.com/support-vector-machines-for-machine learning/
8. http://skyserver.sdss.org/dr2/en/proj/advanced/color/sdssfilters.asp
9. https://www.celestron.com/blogs/knowledgebase/what-are-ra-and-dec
10. Data Release 2 of S-PLUS: Accurate template-fitting based photometry covering 1000 deg2 in 12 optical filters.
11. miniJPAS survey: star-galaxy classification using machine learning.