Stellar Classification

Anthony Obrzut
Ai Yang
Zibo Shang

Project Description and Goal

Project description: Classification of stellar based on their spectral characteristics.

Project importance: The classification scheme of galaxies, quasars, and stars is one of the most fundamental in astronomy. Understanding their distribution helps to understand how our universe is made up.

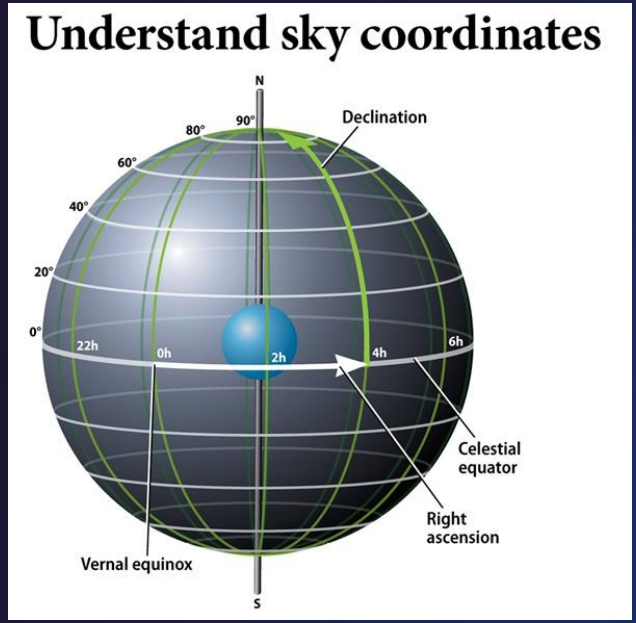
Dataset description: The data consists of 100,000 observations of space taken by the SDSS(Sloan Digital Sky Survey). Every observation is described by 17 feature columns.

Response variable: Class - either a star, galaxy or quasar.

Goal: Correctly classificate stars, galaxies, and quasars based on their spectral characteristics.

Description Of Variables

Class	Response variable: either star, galaxy or quasar
alpha	Right Ascension angle
delta	Declination angle
i	Near Infrared filter in the photometric system
z	Infrared filter in the photometric system



u	Ultraviolet filter in the photometric system
g	Green filter in the photometric system
r	Red filter in the photometric system
redshift	Redshift value based on the increase in wavelength
plate	Plate ID, identifies each plate in SDSS
MJD	Modified Julian Date, used to indicate when a given piece of SDSS data was taken

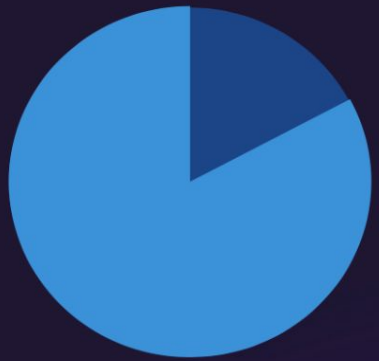
Run_ID	Run Number used to identify the specific scan
Rerun_ID	Rerun Number to specify how the image was processed
cam_col	Camera column to identify the scanline within the run
field_ID	Field number to identify each field
spec_obj_ID	Unique ID used for optical spectroscopic objects
fiber_ID	fiber ID that identifies the fiber that pointed the light at the focal plane in each observation
obj_ID	Object Identifier, the unique value that identifies the object in the image.

Data Split

100,000 Obs w/o NAs

90%

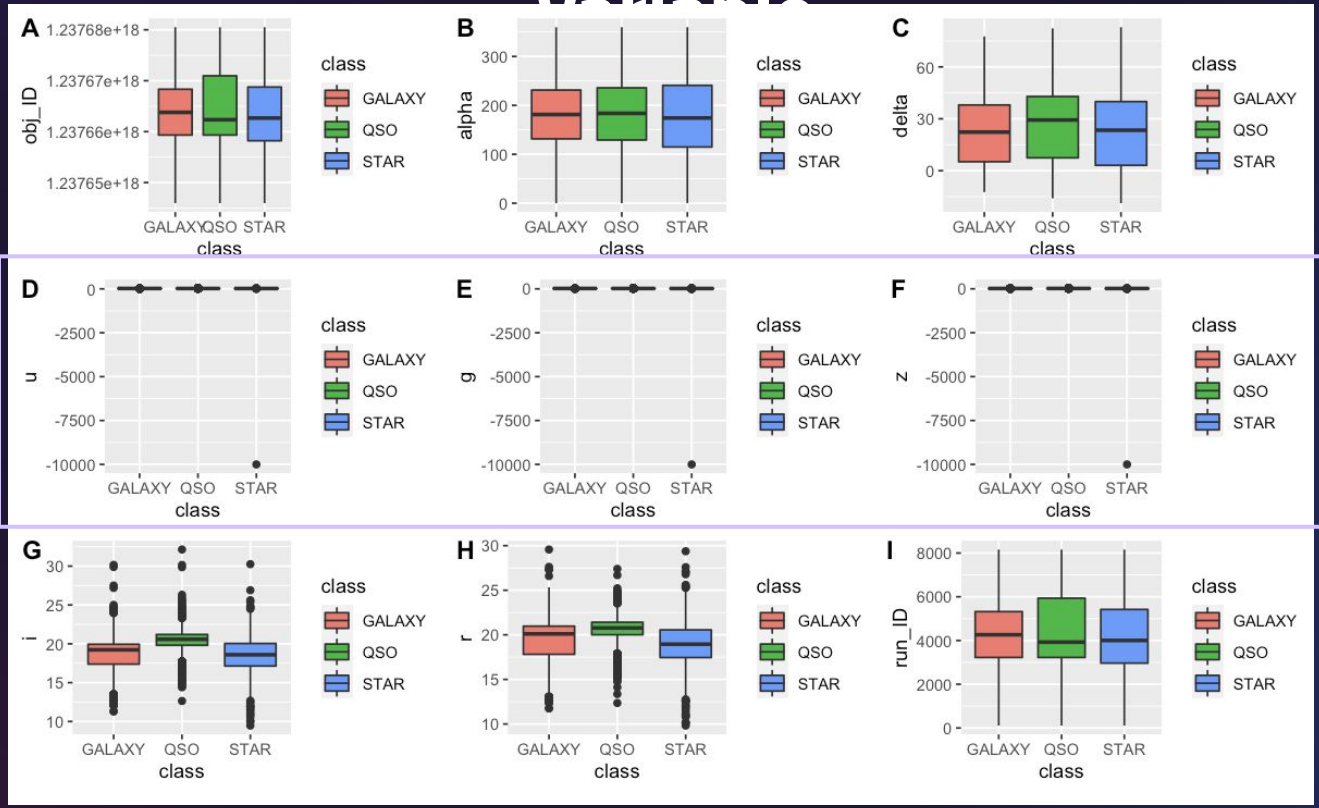
Training:
90,000 data:
6 -fold CV:
15,000 in each fold



10%

Holdout:
10,000 data

Boxplot of individual explanatory variable



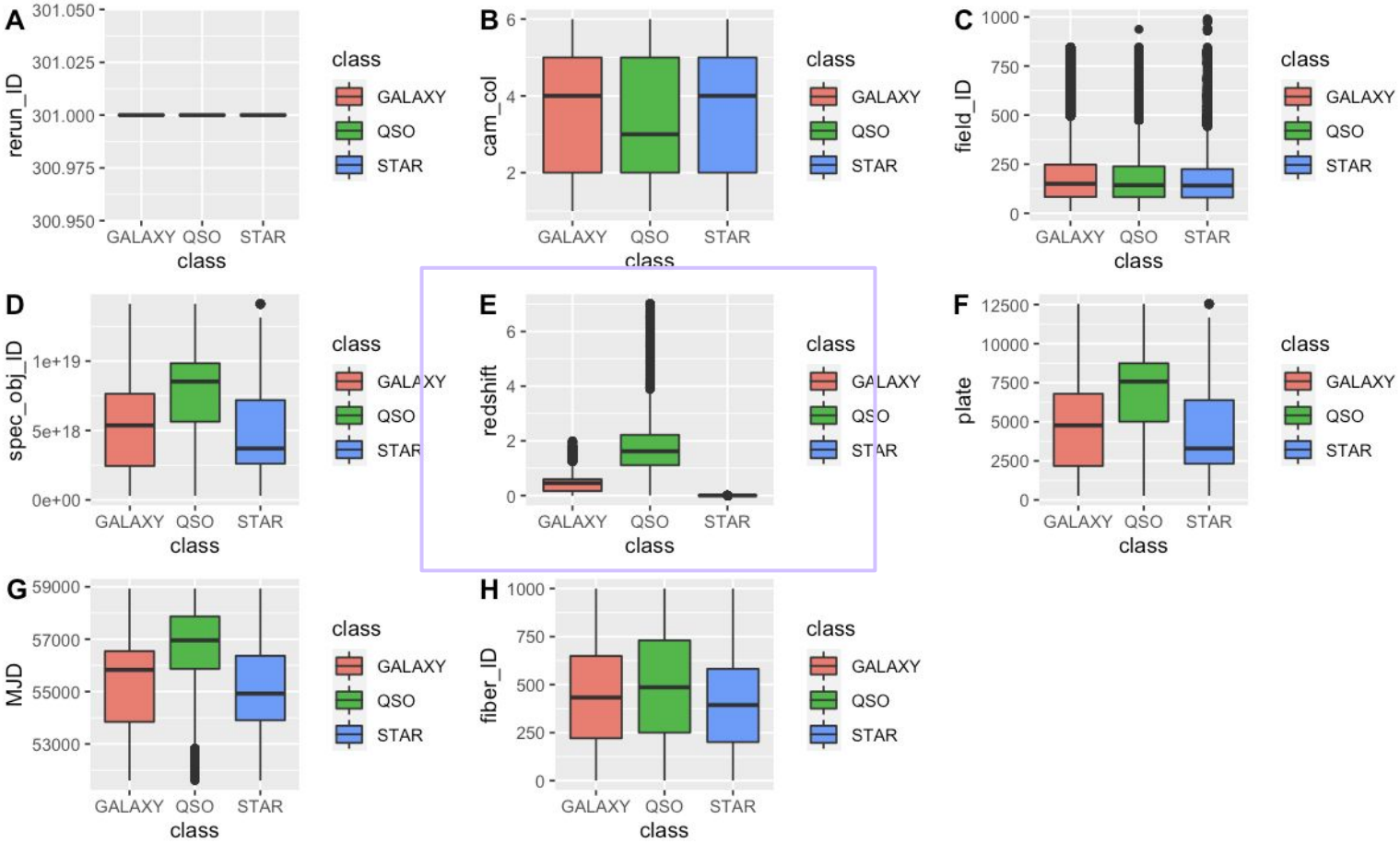
u,g and z: star

```
star[79544,]
```

A data.frame: 1 × 18

	obj_ID	alpha	delta	u	g	r	i	z	run_ID	rerun_ID	cam_col	field_ID	spec_obj_ID	class	redshift
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<int>	<int>	<dbl>	<fct>	<dbl>
79544	1.237649e+18	224.0065	-0.6243039	-9999	-9999	18.1656	18.01675	-9999	752	301	2	537	3.731277e+18	STAR	8.934163e-05

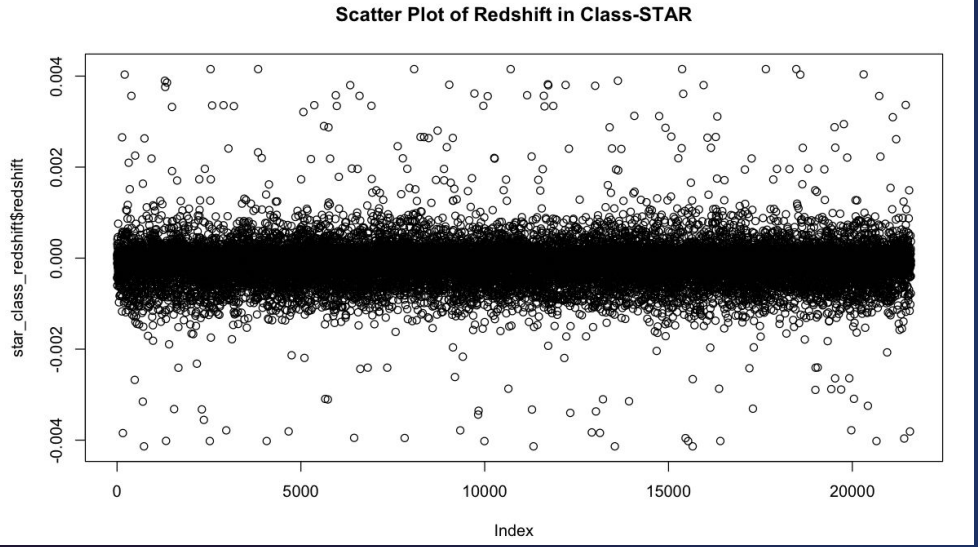
The extreme values of u, g and z come from the same observation (index = 79544). Therefore, we would like to remove this observation to get rid of this extreme value.


A
rerun_ID
301.050
301.025
301.000
300.975
300.950
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
B
cam_col
6
4
2
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
C
field_ID
1000
750
500
250
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
D
spec_obj_ID
1e+19
5e+18
0e+00
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
E
redshift
6
4
2
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
F
plate
12500
10000
7500
5000
2500
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
G
MJD
59000
57000
55000
53000
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
H
fiber_ID
1000
750
500
250
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR

Redshift: star

class	n	min	median	max	sd
<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
GALAXY	59445	-0.0099706670	0.4562742000	1.995524000	0.2648576059
QSO	18961	0.0004606231	1.6172320000	7.011245000	0.9139542763
STAR	21594	-0.0041360780	-0.0000761506	0.004153254	0.0004651542

The redshift data for star is centered around 0, with a small sd around 0.0005.



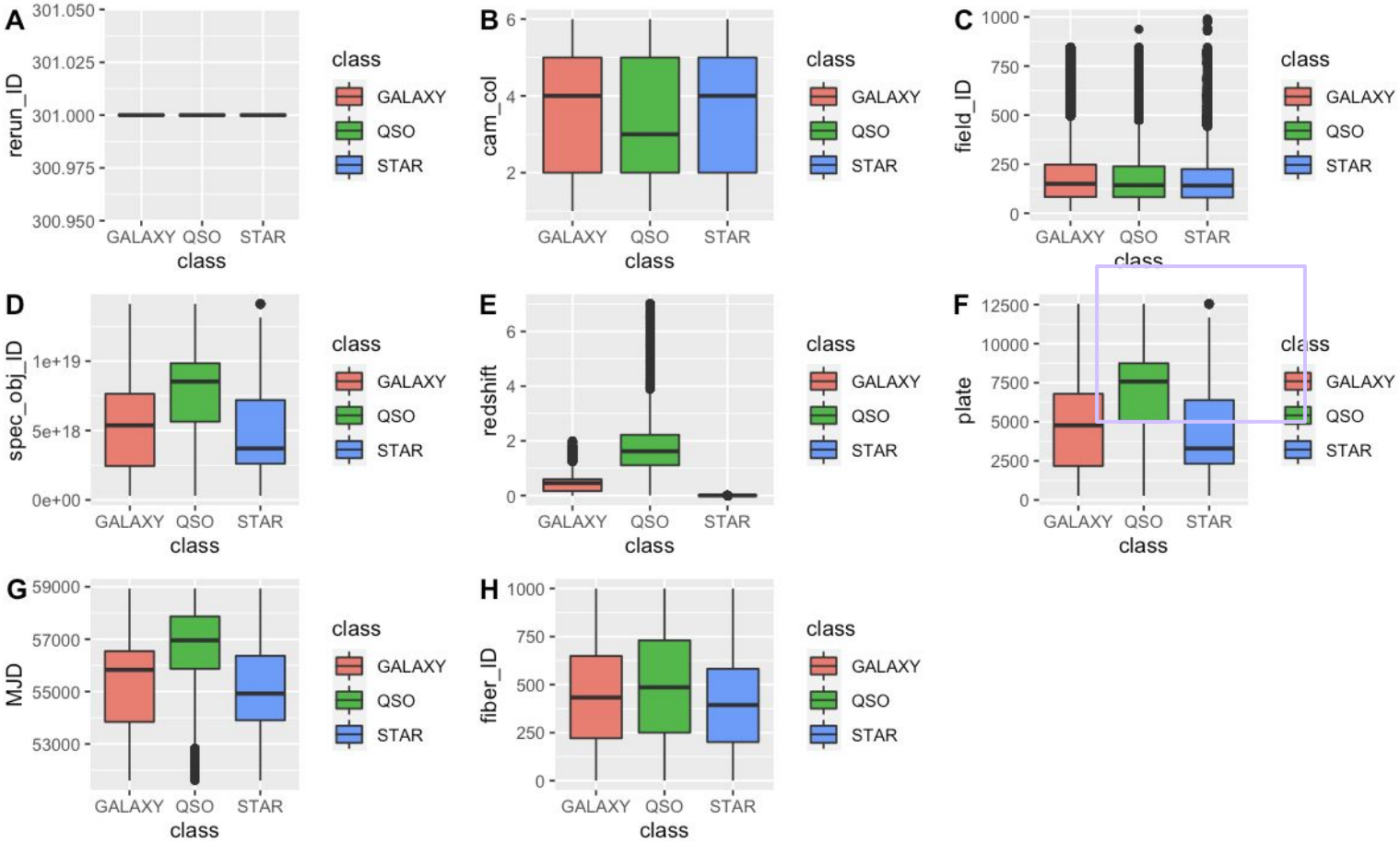
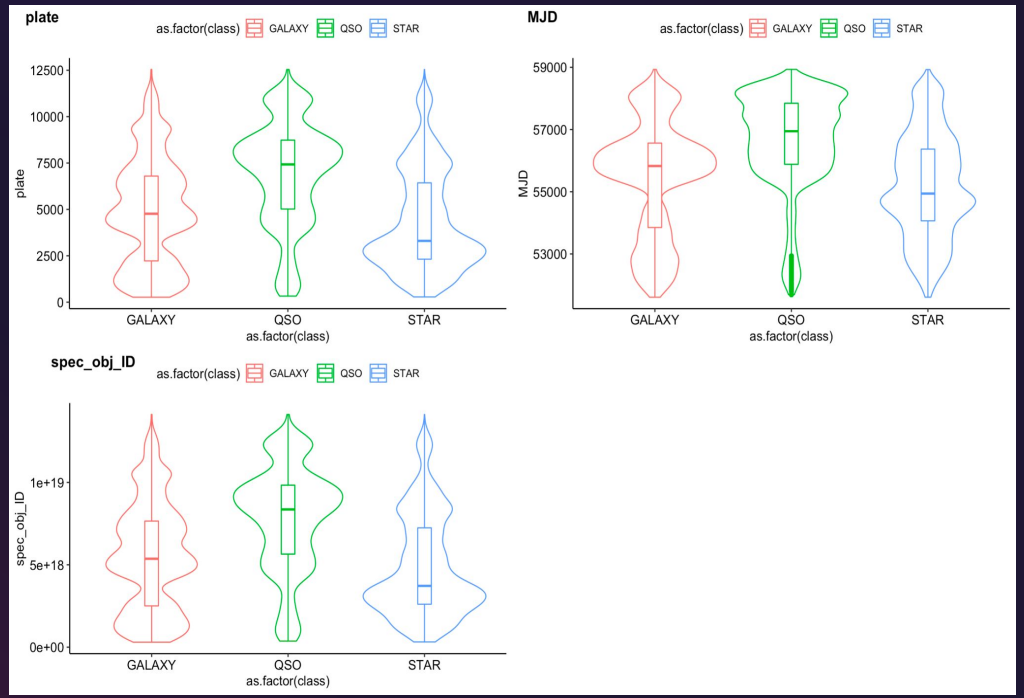

A
rerun_ID
301.050
301.025
301.000
300.975
300.950
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
B
cam_col
6
4
2
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
C
field_ID
1000
750
500
250
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
D
spec_obj_ID
1e+19
5e+18
0e+00
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
E
redshift
6
4
2
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
F
plate
12500
10000
7500
5000
2500
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
G
MJD
59000
57000
55000
53000
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR
H
fiber_ID
1000
750
500
250
0
GALAXY QSO STAR
class
class
GALAXY
QSO
STAR

Plate: star

class	n	plate_unique_value	min	min_count	median	median_count	max	max_count
<fct>	<int>	<int>	<int>	<int>	<dbl>	<int>	<int>	<int>
GALAXY	59445	5557	266	14	4770	22	12547	13
QSO	18961	4165	267	12	7574	41	12547	13
STAR	21594	4617	267	12	3296	24	12547	13

The maximum value of star for plate is equal to the maximum value of galaxy and QSO, so we should keep this extreme value in our dataset.

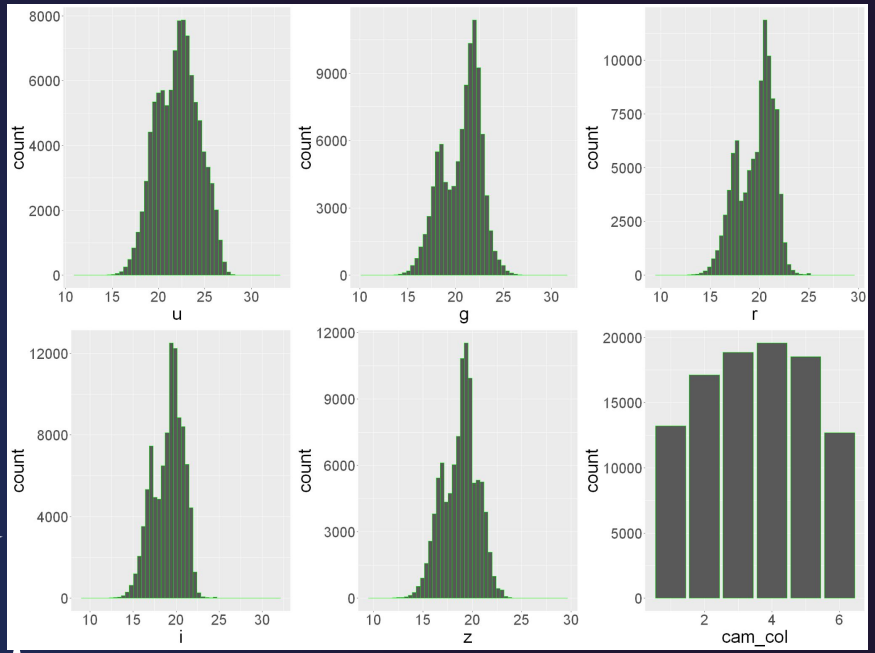
spec_obj_ID, MJD and Plate



- The violin plot of both **plate** and **spec_obj_ID** are identical to each other
- The violin plot of **MJD** is different from the other two, and is different among the three classes.

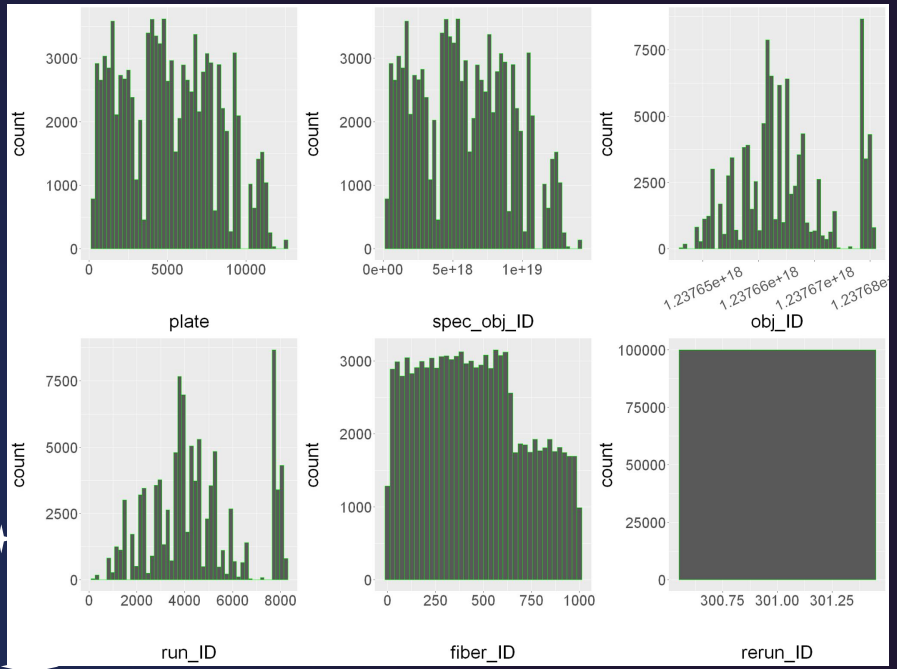
Therefore, we would like to remove spec_obj_ID and keep plate and MJD for future model construction.

Distributions of Our Variables



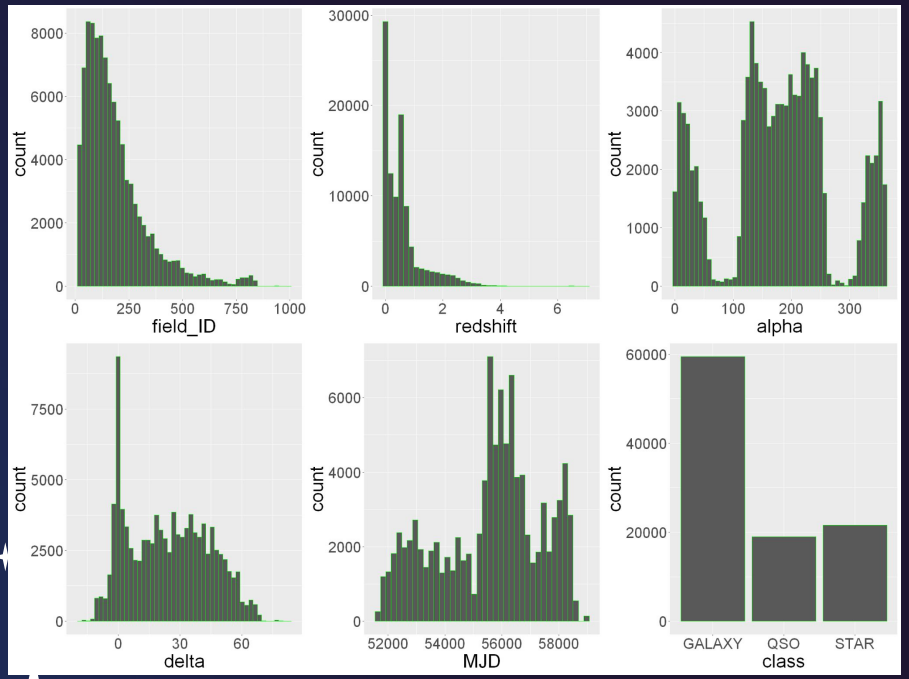
- u, g, r, i, and z all appear to be somewhat normally distributed, and somewhat bimodal.
- This could suggest possible collinearity among these variables.
- cam_col contains somewhat the same number of values for each category, although 3, 4, and 5 appear to have the most values.

Distributions of Our Variables



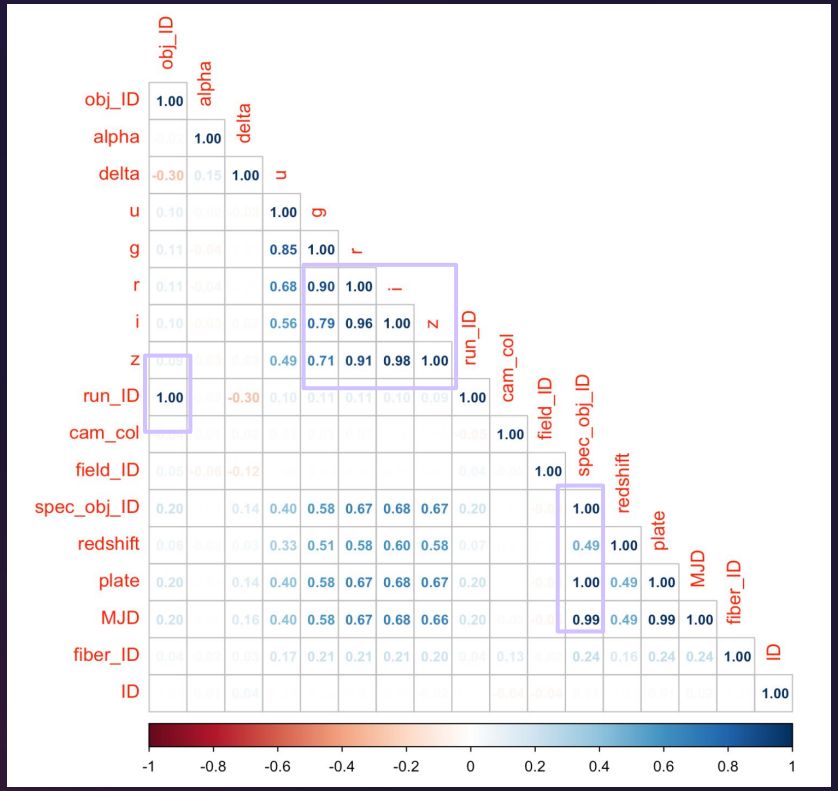
- plate and spec_obj_ID appear to have essentially the same distribution. However, they are slightly different
- Same with obj_ID and run_ID
- This suggests that we should investigate whether plate and spec_obj_ID are correlated, and whether obj_ID and run_ID are correlated
- fiber_ID appears to have two plateaus.

Distributions of Our Variables



- Both field_ID and redshift appear to be right-skewed
- alpha appears to possibly be trimodal, same with MJD, but to a much lesser extent
- delta appears to be somewhat normally distributed, except at about 0, which has a much higher frequency than the other possible values in the distribution
- Our response variable, class, appears to have many objects classified as GALAXY. QSO and STAR have similar relative frequencies.

Correlation Matrix



- **Run_ID** is highly correlated to **obj_ID**;
- Correlation between the filters (u and g, g and r, r and i, i and z) are close to 1.
- **Plate** is highly correlated to **MJD** and **spec_obj_ID**;
- Most of the explanatory variables are **positively correlated** to each other.

Transformation and Wrangling

Table I. Transformation Decision		
Variable Type	Name	Decision
Categorical	Class*	Covert Strings into numeric response
Numeric(dbl)	alpha	Standardization
Numeric(dbl)	delta	Standardization
Numeric(dbl)	i	Standardization
Numeric(dbl)	u	Standardization
Numeric(dbl)	g	DROP, high correlation
Numeric(dbl)	r	DROP, high correlation
Numeric(dbl)	z	DROP, high correlation
Numeric(int)	obj_ID	DROP, high correlation
Numeric(int)	cam_col	trans to categorical by factor
Numeric(int)	rerun_ID	DROP, same for all
Numeric(int)	spec_obj _ID	DROP, meaningless by definition
Numeric(int)	field_ID	Standardization and Binning
Numeric(dbl)	redshift	Standardization and Binning
Numeric(int)	run_ID	Standardization
Numeric(int)	plate	Standardization
Numeric(int)	MJD	Standardization
Numeric(int)	fiber_ID	Standardization

Model

Three types of response variable

Unbalance response variable

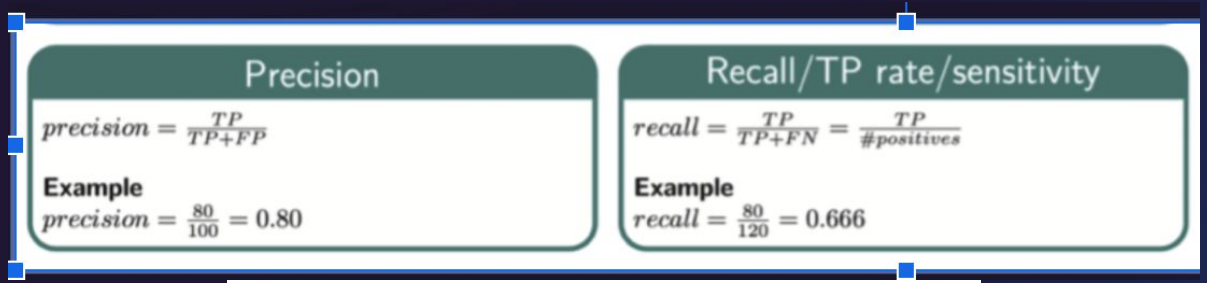
Multinomial logistic regression

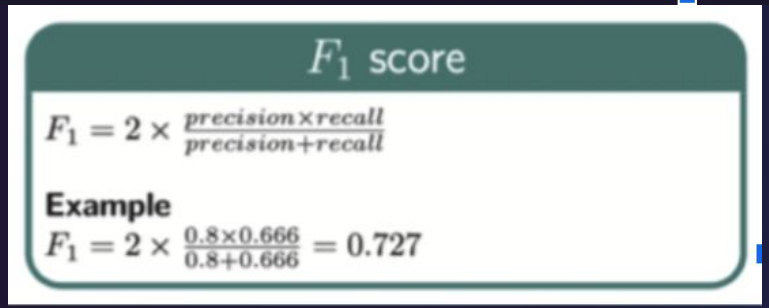
Naive bayes

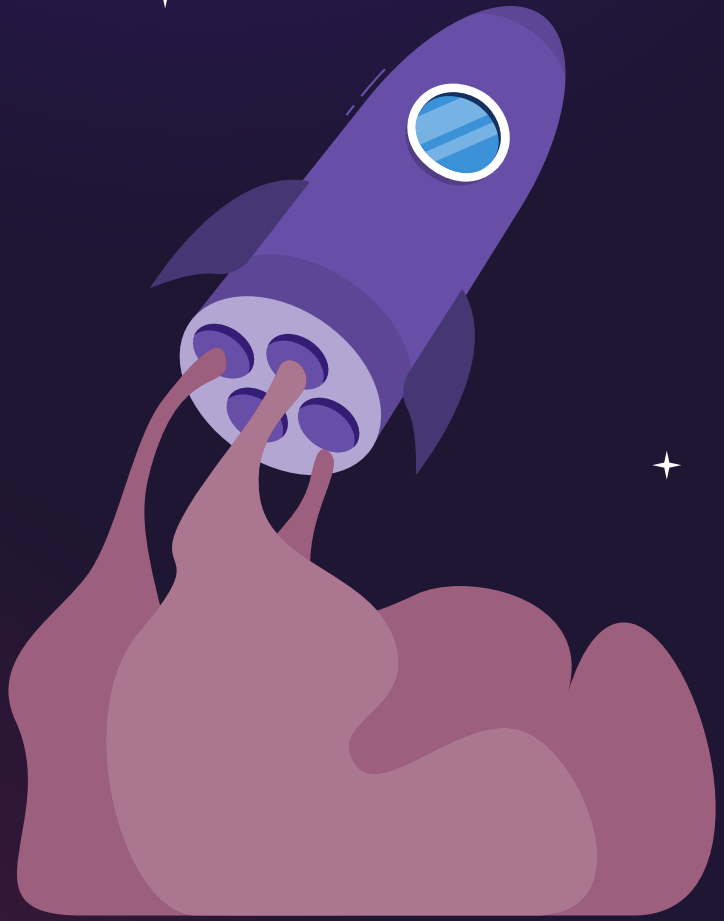
Random forest/XGBoost

Metric Selection

F1 Score
80% Predictive Interval







THANKS!