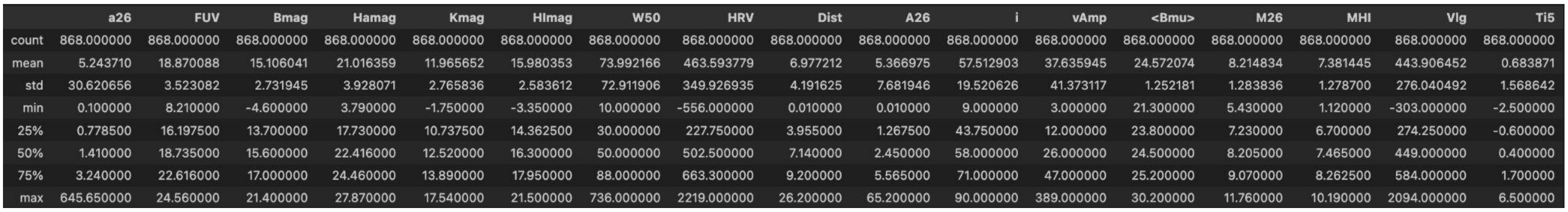
Milestone 3

Summary of the Data:

The dataset we are using is a galaxy catalog, which contains close to 1000 galaxies, with features such as type, and other photometric properties. Our response variables will be Morphological Type (TT), and Dwarf Galaxy Morphology (Mcl). Our predictors and summary statistics are as follows:

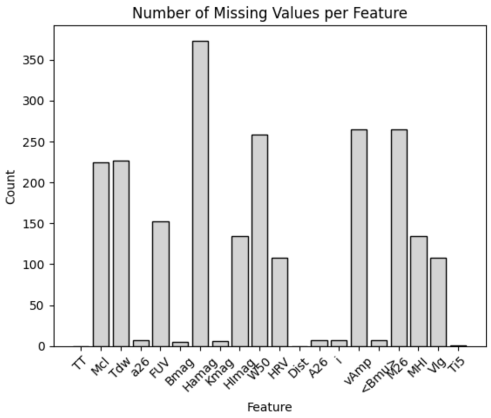
- **a26**: Major angular diameter
- **FUV**: Far UV
- **Bmag**: Apparent integral b-band magnitude
- **Hamag**: Integral H-alpha line emission magnitude
- **Kmag**: K-band magnitude
- **HImag**: HI 21cm line magnitude
- **W50**: HI Line width at 50% level from maximum
- **Tdw**: Dwarf galaxy surface brightness morphology
- **HRV**: Heliocentric radial velocity
- **Dist**: linear distance to galaxy from the sun

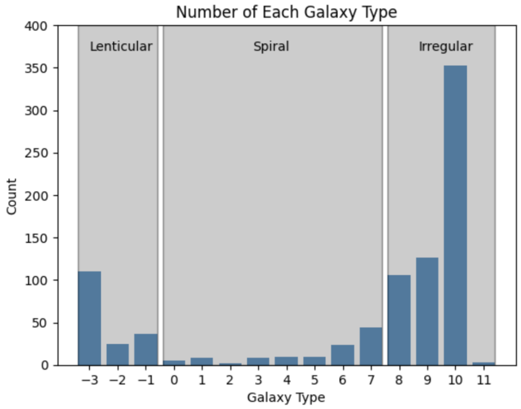
- **A26**: major linear diameter
- **i**: inclination of galaxy from the face on
- **vAmp**: amplitude of rotational velocity
- **Bmu**: average b-band surface brightness
- **M26**: log-mass with Holmberg radius
- **MHI**: log-Hydrogen mass
- **Vlg**: local group radial velocity
- **Ti5**: tidal index
- **SimbadName**: designation understandable by the Simbad astronomical database



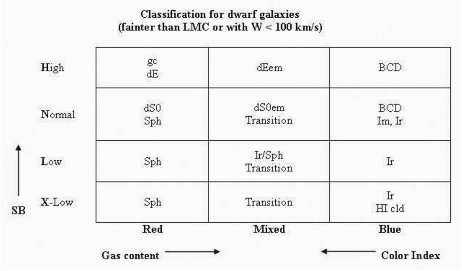
	a26	FUV	Bmag	Hamag	Kmag	HImag	W50	HRV	Dist	A26	i	vAmp	<Bmu>	M26	MHI	Vlg	Ti5
count	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000	868.000000
mean	5.243710	18.870088	15.106041	21.016359	11.965652	15.980353	73.992166	463.593779	6.977212	5.366975	57.512903	37.635945	24.572074	8.214834	7.381445	443.906452	0.683871
std	30.620656	3.523082	2.731945	3.928071	2.765836	2.583612	72.911906	349.926935	4.191625	7.681946	19.520626	41.373117	1.252181	1.283836	1.278700	276.040492	1.568642
min	0.100000	8.210000	-4.600000	3.790000	-1.750000	-3.350000	10.000000	-556.000000	0.010000	0.010000	9.000000	3.000000	21.300000	5.430000	1.120000	-303.000000	-2.500000
25%	0.778500	16.197500	13.700000	17.730000	10.737500	14.362500	30.000000	227.750000	3.955000	1.267500	43.750000	12.000000	23.800000	7.230000	6.700000	274.250000	-0.600000
50%	1.410000	18.735000	15.600000	22.416000	12.520000	16.300000	50.000000	502.500000	7.140000	2.450000	58.000000	26.000000	24.500000	8.205000	7.465000	449.000000	0.400000
75%	3.240000	22.616000	17.000000	24.460000	13.890000	17.950000	88.000000	663.300000	9.200000	5.565000	71.000000	47.000000	25.200000	9.070000	8.262500	584.000000	1.700000
max	645.650000	24.560000	21.400000	27.870000	17.540000	21.500000	736.000000	2219.000000	26.200000	65.200000	90.000000	389.000000	30.200000	11.760000	10.190000	2094.000000	6.500000

Across the features we use, there are many rows with missing values in one feature or another. All galaxies have a type and a distance though, and most have an angular diameter, a B magnitude, a K magnitude, an inclination, and a Tidal Index. The initial Galaxy Type Distribution goes as follows, where each number corresponds to the Morphological type from Vaucouleurs et al. (1991). The corresponding Hubble Type is plotted on top. Of our total galaxy sample, 172 of them are Lenticular, 280 are Spiral, 126 are a mix of spiral/irregular, and 350 are Irregular.





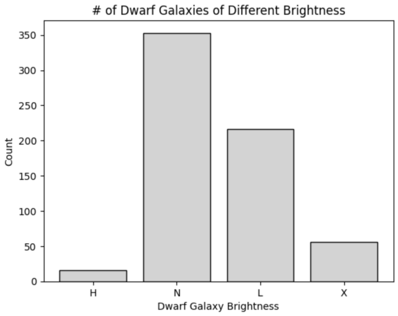
However, this does not take into account that most (75%) of our galaxies are actually Dwarf Galaxies. Within dwarf galaxies, there is a separate classification scheme of brightness and color. Most of our dwarf galaxies are Normal or Low Brightness, and Blue.

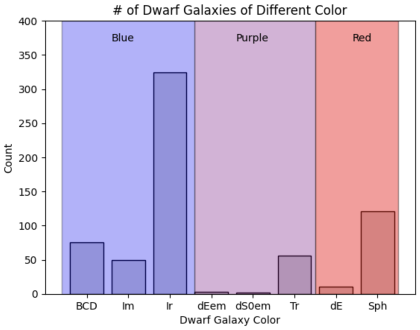


Classification for dwarf galaxies
(fainter than LMC or with W < 100 km/s)

High	gc dE	dEem	BCD
Normal	dS0 Sph	dS0em Transition	BCD Im, Ir
↑ Low	Sph	Ir/Sph Transition	Ir
SB X-Low	Sph	Transition	Ir HI cld
	Red	Mixed	Blue

Gas content → ← Color Index





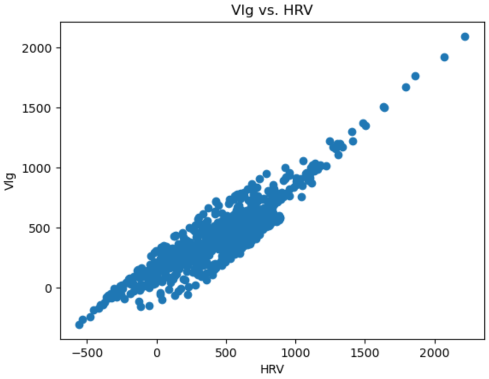
Data Description:

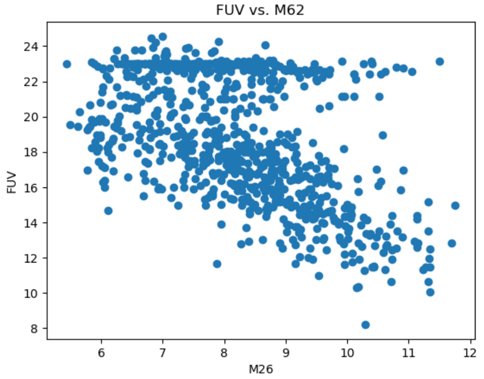
We initially were trying to use the Galaxy10 DECals dataset, which contains galaxy images, positions, and types, and crossmatch that with a database of galaxy positions and features (such as Star Formation Rate and Mass). After consulting with several experts in the field, we decided to use an Astronomical Catalog software (TOPCAT). We successfully installed and used the software, but none of the virtual databases had what we wanted. We decided to pivot our database and opt for something already consolidated and with much nearer galaxies. The data source we use is the Updated Nearby Galaxy Catalog, which was compiled in 2013 and contains 869 nearby galaxies that are within 11 Mpc, or have small velocities relative to the Local Group (<~600 km/s). The sample originated from a list of 179 nearby galaxies compiled by Kraan-Korteweg & Tammann (1979), and which was further added to by large all-sky surveys like the Sloan Digital Sky Survey. Distances to each galaxy were found using a variety of astronomical distance tools (such as Cepheids and Supernovae), as well as the Hubble velocity–distance relation of H_0= 73 km/s/Mpc (which says that further away galaxies are moving away faster, because the universe is expanding).

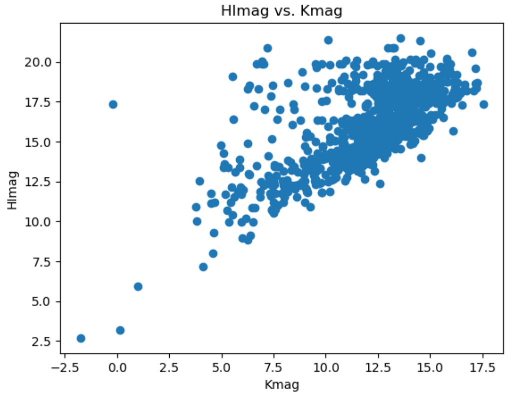
On our end, the data collection process began by finding the right catalog for our needs, which took some time. We settled upon this one after a thorough search of many different galaxy catalogs. Once we had decided on it though, we were able to download it fairly easily from **Vizier,** which holds a complete library of published astronomical catalogs. We loaded the **.fits** file into a jupyter notebook using **pandas.** The initial steps taken to understand the data included reading through each of the features (some of which are quite technical, and which we were not familiar with), and plotting some histograms of each as well as scatterplots of the correlations between them. Not much data cleaning was necessary, although several galaxies had missing values in one or more features. We dropped certain features from the initial database which we found to be either repetitive, useless, or for which the majority of our galaxies were missing (such as the upper/lower limit flags on certain magnitudes). We further drop rows of our database if 50% or more of the predictors are missing. For the other missing values, we impute them with KNN imputation (k=5).

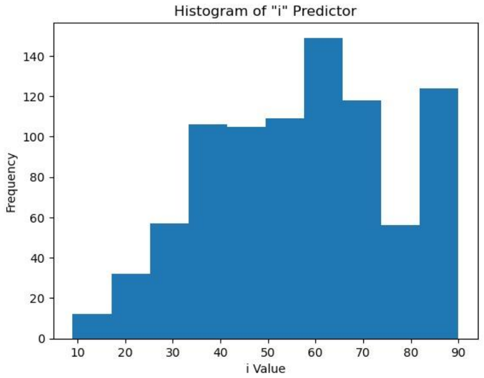
EDA and Noteworthy Findings:

We began our EDA by making pairplots of our data to gain insight into the relationships between predictors and how they may be clustered by galaxy type. We found a couple of interesting relationships which we outline below:


Vlg vs. HRV

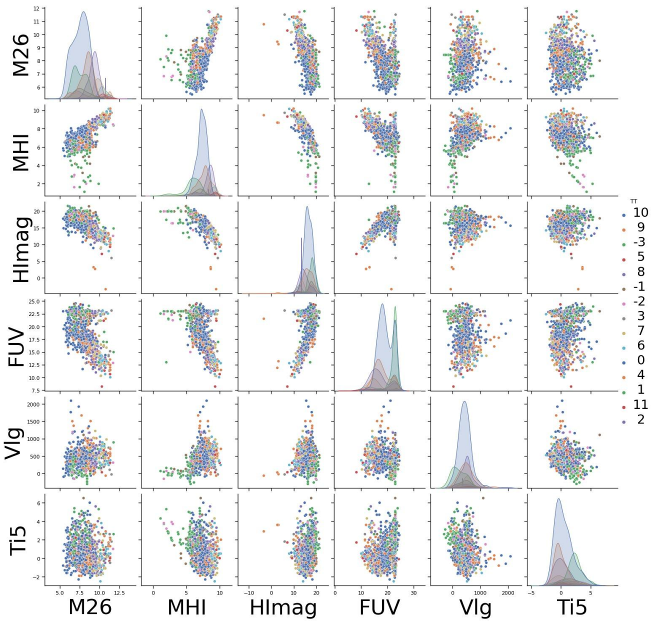

FUV vs. M62


HImag vs. Kmag

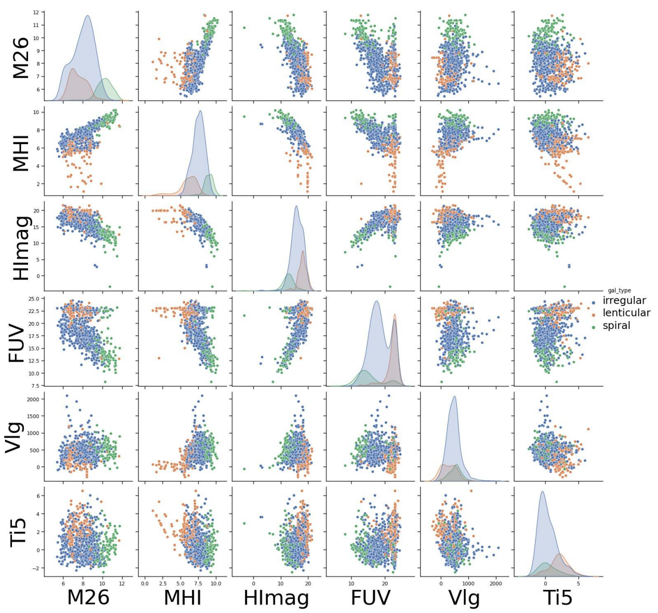

Histogram of "i" Predictor

1. Top left: HRV (Heliocentric radial velocity) and VLG (Local group radial velocity) are highly correlated and basically collinear. This is expected because they are both radial velocities. We will choose to ignore HRV because the velocity with respect to the local group of the galaxy is far more intrinsic to the galaxy that its velocity with respect to the Sun.
2. Top middle: Many magnitude values are correlated, which we expect since it refers to the brightness of the galaxy, just in different wavelengths
3. Top right: FUV (far UV magnitude) and M26 (log(mass) within Holmberg radius) seem negatively correlated, which makes sense because we expect more massive galaxies to be brighter, and lower magnitude corresponds to greater luminosity.
4. Bottom right: ‘i’ (inclination of the galaxy from the edge on) favors large values. This we expect because galaxies should be more difficult to identify/observe if we find them edge on.
5. Perhaps most importantly, we made a pariplot of all the predictors that we found potentially be physically important ('M26', 'MHI', 'HImag', 'FUV', 'Vlg', 'Ti5') colored by the morphology T-type, overarching Hubble class, and dwarf morphology type. We find that all of these are, at least visually, somewhat clustered by type. Overarching Hubble class is definitely the most clustered (see below). Below we also include T-type.

T-Type:



Hubble Class:



Project Question:

After gathering the data, and starting our EDA we have come up with our initial question / motivation to guide our project:

> **What predictors are most important for predicting the type of galaxy? (galaxy type defined by Hubble Class, T-type, and/or dwarf types)**

To answer this question, we will attempt to train a model(s) to predict galaxy type with our set of predictors, and explore the specific relationships between the predictors and type.

We are also going to examine the relationship between different features to see if we can recover known astronomical relationships from our dataset (like the Hubble Law and Fisher-Tully Relation).

Baseline Model or Implementation Plan:

We will attempt to train and classify our data using kNN classification, multiple logistic regression, and random forests with gradient boosting. We will also explore what features might be of poor predictive power in our dataset, and see if we can reduce the number through techniques like PCA. If these methods are not robust enough to effectively classify our dataset, we will train a neural network and see if it does better. We will refine our models using k-fold cross-validation, and evaluate them using binary cross entropy. We will summarize these results in a confusion matrix to see where each model struggles, beyond just seeing classification errors.