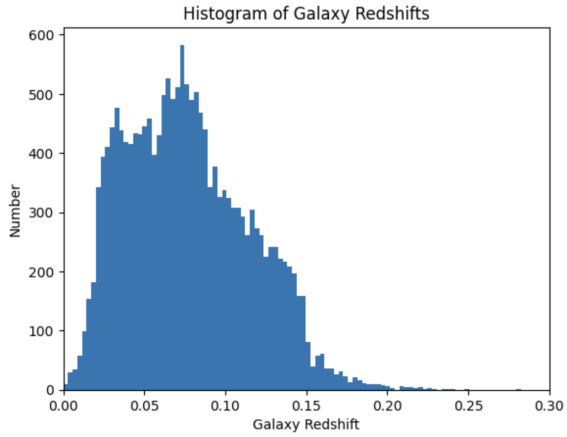
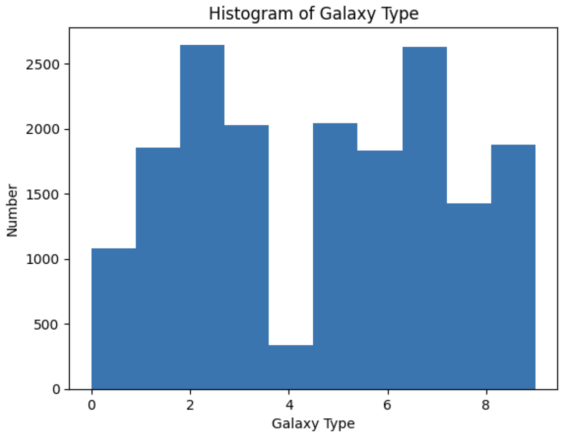
Our data consists of two primary datasets, the Galaxy10 DECals dataset and a catalog of redshifts, stellar masses, and star formation rates compiled by Zou et al. (2019). The end goal of our project is to use methods from class to predict galaxy type given the galaxy images from the Galaxy10 DECals dataset; time permitting, we will also use the images to predict mass, SFR, photometric redshift, age, and other properties of the DECals galaxies inferred in the Zou et al. dataset. We plan to explore how methods such as principal component analysis, k-nearest neighbor regression, and random forests can be used to group galaxies based on their images, and how their efficacy compares to methods like CNNs

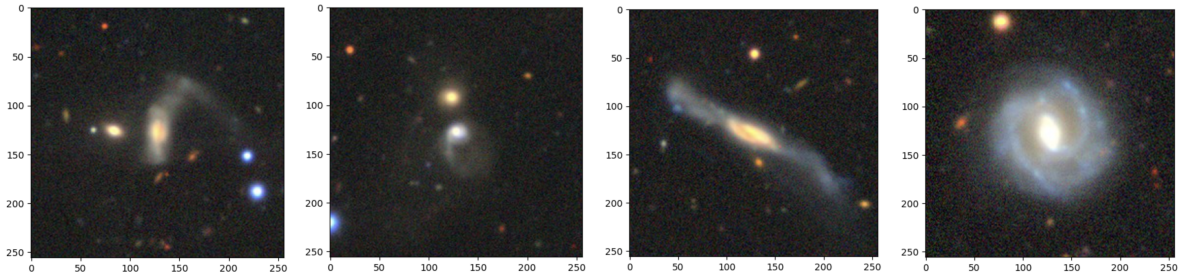
Dataset 1: Galaxy10 DECals

The Galaxy10 DECals dataset is well compiled, and has no major issues. We downloaded the dataset as a **.h5** file (~2.5Gb) and loaded it into python. There are 17,736 galaxies in this table, which also records the galaxy RA, DEC, redshift, and morphological type. The RA, DEC values represent the galaxy's position in the sky, the redshift corresponds to a measure of distance, and the morphological type represents what kind of galaxy it is (Barred, Merging, Spiral, Smooth…). There are 10 galaxy types, and a fairly even distribution of all of them except for Class 4 (for which there are <400).





The galaxy images also all seem to be in good order, and are easily accessible. They are RGB images of 256x256 pixels.



Overall, no data is missing from this dataset. None of the predictors need to be scaled either, as we are primarily using the galaxy type from this dataset as our response variable. One issue

could be the imbalance in Galaxy Type, specifically the low number of galaxies in Class 4. We will need to be aware of this when training any models. One potential remedy could be to undersample the rest of the classes, or oversample Class 4.

Dataset 2: Catalog of Photometric Redshifts and Stellar Masses for Galaxies from the DESI Legacy Imaging Surveys

This catalog consists of galaxy properties for galaxies imaged by the Dark Energy Spectroscopic Instrument (DESI). The imaging footprint of DESI covers an area of 14,000 deg^2 (a massive portion of the sky in astronomy's standards), so there is a *ton* of data and the file sizes ended up being much larger than we expected based on our preliminary analysis. Our plan with this data was to match the galaxies to the galaxies in DECals using a cone search (essentially pair the coordinates with the smallest angular separations). We downloaded the file photoz.fits.gz from this link. Unzipping this yields a 78 GB fits file containing the properties of 303,379,640 galaxies. This is obviously a *massive* file for our computers to handle which has caused us immense difficulties. We attempted to filter the file by only including galaxies in the 0-60 degree declination range, and our computers spent close to an hour before exceeding the memory limit and crashing. We used multithreading to obtain a mask based on the declination values, and this was successful (after 15 minutes of run time), but applying the mask to any array still caused our code to crash. See the multithreading code below:

```
from multiprocess import Pool, cpu_count

# Where the majority zoo data is
min_dec = 0
max_dec = 60

# Functions for parallel processing
def get_mask(data_chunk):
    """Function to apply mask on a chunk of data."""
    print('Getting mask for chunk')
    mask = (data_chunk >= min_dec) & (data_chunk <= max_dec)
    return mask

def parallel_mask(data, chunks):
    """Function to apply mask in parallel."""
    with Pool(processes=cpu_count() - 2) as pool:
        print(f'Using {cpu_count() - 2} cores')
    result = pool.map(get_mask, np.array_split(data, chunks - 2))
    return np.concatenate(result)
```

Going forward, we will consult people with field experience to find the best way to match the galaxies in our data. We are considering using different galaxy property databases or uploading the file to the FASRC cluster computer to clean it with more computational power. Please let us know what you recommend towards this end. Then, we can begin predicting the properties using the galaxy images!