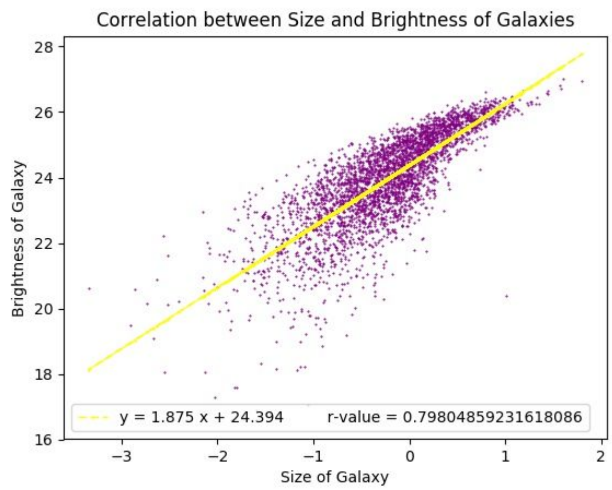
Analyzing Galactic Environments
by: Elizabeth Lin and Lily Jiang

1) Summary of research questions and results.

1. How does the brightness of galaxies affect the size of the galaxy?

We will be examining the brightness in a sequence (420 nm, UV to 915 nm, far red) given in linear variables and the total R magnitude which is the measurement of brightness (the smaller R, the brighter). We will compare the brightness of each galaxy and the ApDRmag (difference between total and aperture magnitude in R band), which is the measure of the galaxy size, to compute a value to hypothesize the relationship between brightness and galaxy size. Our end result is to examine how brighter galaxies can impact the chemical composition of the solar system and finding life on potential new planets.



Results: The result of our graph show that there is a positive correlation between the size and brightness of a galaxy. The greater the size of the galaxy (according to magnitude of the red band), the brighter the galaxy. We also calculated the p-value between the size and brightness which we obtained was 0.798048592316. This is a high p-value so it shows that both elements have strong correlation (Pearson correlation is the measure of the linear correlation between two variables on a scale of -1, negative correlation to +1, positive correlation).

2. Categorize galaxies based on their redshift value to analyze the movement/speed.

Edwin Hubble in 1935 discovered that the universe is not held at a constant value and is fact expanding rapidly. The space between galaxies is growing and galaxies are moving away from Earth at a tremendous value. To analyze the groups of galaxies traveling at different speeds, we need to look at the redshift value.The faster the galaxies is moving, the more red shifted its light is and the older the star will be. We can find the first quartile, median, and third quartile of the galaxy's redshift value and group them accordingly. This will show what is the most common speed galaxies are traveling away from the Earth. Redshift value can also tell us whether we have more young or old stars in our data set compared to the sun.

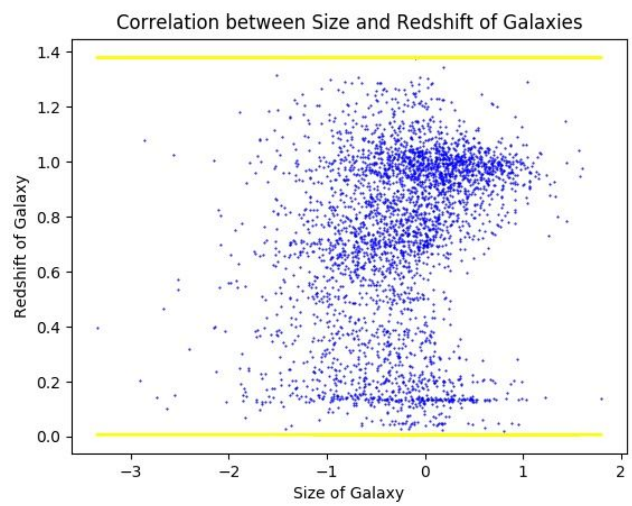
Results: A majority of the galaxies are moving in the fastest category. Because 39.83% of galaxies are moving at a faster redshift value rate, we can conclude that many of the galaxies that we can detect are farther away from the Earth and more likely to be older than our sun.

729 galaxies have a speed between 0 and 0.405.
1008 galaxies have a speed between 0.405 and 0.81.
346 galaxies have a speed between 0.81 and 0.905.

1379 galaxies have a speed between 0.905 and 1.

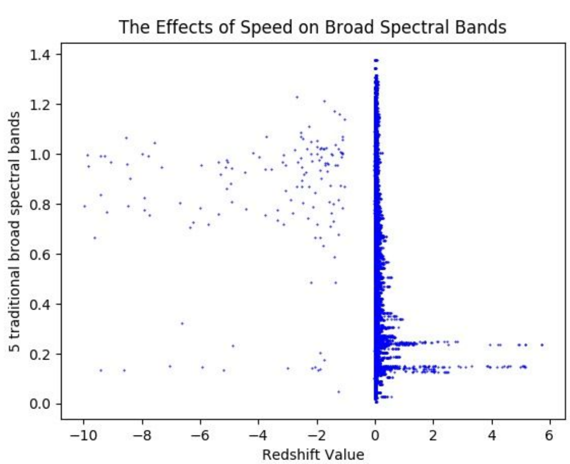
21.0571923744 % of galaxies are moving at a very slow pace
29.116117851 % of galaxies are moving at a slow pace
9.99422299249 % of galaxies are moving at a fast pace
39.8324667822 % of galaxies are moving at a very fast pace

3. How does the size of the bands of a galaxy affect the red shift values of a galaxy?
Looking at previous data about the red shift value is, we can better compare the rate of movement of each galaxy with their absolute magnitude of their bands. The bands are a clear indicator of the galaxy size which can provide a better understanding of galaxies characteristics. The greater their band there is can potentially affect the overall characteristics between its growth and brightness. Using the previous data from question two, we can better understand the correlation between galaxy size and speed. This information will help astronomical scientists develop insight on varying characteristics of galaxies and see if there is a pattern in their mechanism and physical size.



Results: The size of our galaxy does not seem to have any correlation with the red shift values of the galaxy. As observed from our graph, there is a random scatterplot of values that shows no correlation.

4. Find the effect of redshift on observed brightnesses in 13 bands.
Each of the galaxies in our data set gives off a unique value of brightness and color to to the galaxy's chemical composition. But as the galaxies reset farther away from Earth, their light gets distorted and stretched into longer wavelengths. The photons lose energy as they travel and that is why farther galaxies look more red than closer ones. Researching the observed brightnesses in 13 bands from 420 nm in the ultraviolet to 915 nm in the far red, we will see if the majority of our detected galaxies are redshifted or can still be observed in other color ranges.



Results: We can observe galaxies in different color spectrums but a majority of them lie between 650 - 800 nm for color which equates to red. This indicates most of the observed galaxies are far away from us or is currently moving away.

2) Motivation and background.

Recognizing the universe is actually expanding instead of sagnant might possibly be one of the most profound discoveries ever made in astronomy. Scientist thought that the universe and all the galaxies in it were unchanging and fixed without a beginning or an end. Einstein knew that gravity would attract objects together over time, therefore, there must have been another force that counter acted the force of gravity. He called it the cosmological constant and added it to his theory of general relativity. But when Einstein realized that the universe is not static but actually expanding, he abandoned the constant calling it the '"biggest blunder" of his life. The astronomy community believed that the universe is expanding but would slow down in the future. A startling discovery in recent years provided that Einstein's biggest blunder was actually correct! The universe was not only expanding but speeding up at a rapid pace.

To calculate how fast distant galaxies are moving away from us, we can look at the light emitted by these galaxies. An atom or molecule gives off light at a very few particular wavelengths. When light travels towards the Earth, it loses energy, stretching the light waves. So when a galaxy in the distance is farther away, the light it emits takes more energy to reach us. Hence why the more distant the galaxy, the redder it appears. We can measure the redshift of these distant galaxies and to see for ourselves how fast the universe is expanding. Another quality we can look for in these galaxies are the size and brightness. Using the brightness of the galaxy, we can determine the chemical components of its solar system and potentially predict if there is an area of habitable zone for life.

3) Dataset

URL: http://astrostatistics.psu.edu/datasets/COMBO17.html

The dataset is provided by the Center for Astrostatistics at Penn State University. "The Center serves as a crossroads where researchers at the interfaces between statistics, data analysis, astronomy, space and observational physics collaborate, develop and share methodologies, and together prepare the next generation of researchers."

We will be working with the CASt dataset which contains COMBO17.dat and COMBO17.csv file. We may plan to use DictReader function in order read the file and process the information in the file. The file contains 65 columns about 3,462 classified galaxies with information about the brightness measurement, Nr (object number), Total R magnitude, ApDRmap (difference between total and aperture magnitude in the R band), Mcz and MCzml redshift estimates, and absolute magnitudes of the galaxy in the 10 bands and their measurement errors.

4) Methodology (algorithm or analysis)

For this part of our project, we will be crafting an efficient program that analyzes the data and to do this, we will have to brainstorm the various parameters and outputs in each function. First to extract the data, we wrote a function that takes a filename and a list of column names and returns a list of integers that contains the values in those columns from every row. This is how we extract the numbers from the csv files. Next we wrote a function called get size that takes the difference between the values of two lists and returns the values in a list. This is a helper function that will be implemented to graph other functions. There were also other helper functions written like get_gal_number, which returns list of names of galaxies, and get_redshift, which gets redshift value of all galaxies.

For the first question to graph size correlation to brightness, we wrote a function that takes the x-axis as the size and the y-axis as the brightness. Next we used linear regression to calculate the equation of the line and found that there is a positive correlation. For question 3, we implemented the same strategy with size and the redshift value but found that it does not have any correlation to each other. For question 2, we used a function to categorize the galaxies according to their redshift value. We then translated the groups into percentage values to see which redshift value group had the most amount of galaxies in it. Lastly, question 4 used the same methods of plotting to compare the visible color spectrum to redshift value with y as the redshift value and x and the visible bands. With this, it resulted that an overwhelming amount of galaxies appear red.

5) Results:

For our first research question, we were examining what happens to the brightness of a galaxy as a galaxy gets bigger (measured in R bands). We computed that the relationship was linear; the greater the galaxy in size, the greater the brightness. This was similar to our hypothesis because a galaxy taking up a larger area would emit a greater light due to it having more energy than smaller galaxies.

For our second research question, we first calculated the first quartile, median, and third quartile of the galaxy's redshift value. They are 0.405, 0.81, and 0.905 respectively. These numbers then became the parameters of the category. From this data we can conclude that a majority of the galaxies, 39.83%, are traveling in the fastest category and moving away from Earth at top speeds. There were only 9.99% of galaxies moving at a red shift speed between 0.81 and 0.905, making it the category with the fewest number of galaxies. Because 39.83% of galaxies are moving at a faster redshift value rate, we can conclude that many of the galaxies that we can detect are farther away from the Earth and more likely to be older than our sun.

For our third research question, we looked into what happened to the redshift values as galaxies get bigger. Initially, we hypothesized that this change would allow for greater red shift values because the bigger galaxies are, the faster they can move because they have more kinetic energy to emit in the solar system. However, according to the graphical representation of our plotted graph, there was a scatter of plot that suggested there is no correlation between galaxy size and their redshift values. In fact, we examined the maximum and minimum redshift values did not occur in distinctly different galaxies.

For our fourth research question, a majority of our data lies near or close to 0 and 1 on the redshift value and is within the 650 - 800 nm for color. This proves that most galaxies are moving aways from us and the light is distorting to a more red hue. We can still observe galaxies that are moving at a slower pace in other color categories but the graph can place most confidently in red.

6) Reproducing your results.

1. **Obtaining files**:
   a. Download zip file (galaxies.zip).
   b. Open the file contents in various software (installed).
      i. galaxies.csv in TextEdit
         1. Terminal *cd Downloads/galaxies*
         2. *open -a textedit galaxies.csv*
      ii. main.py in Canopy
      iii. *open -a canopy main.py*

* Make sure on Terminal that you are in your galaxies directory to access these files!

2. **Run files:**
   a. Once main.py is opened, press command+R to run the file
      i. A graph of Brightness vs. Size graph should pop up called: graph-brightness-sizes.jpeg
      ii. A graph of Brightness vs. Redshift graph should pop up called: graph-redshift-sizes.jpeg

To interpret the results, you can look closely at the graph to see what specific correlation (linear or scattered) of the data set.

7) Work Plan Evaluation.

Part One:

We will be grabbing our data from the dataset and determining what our end product or result will be. We will be choosing the brightness, size, and redshift value of each galaxy in the data and will be analyzing the data through computations to answer our research questions. To do this, we will be using the graph function we used for Homework 4 and the DictReader function to convert the csv file we are provided as our dataset. We spent thirty minutes for this portion of the project.

Part Two:

For the first research question, which is investigating the correlation between the brightness and the size of the galaxy, we will be plotting the brightness levels of each galaxy on a graph. We will also create an indexed list of the sizes of each galaxy. Having categorized the brightness levels and sizes, we will utilize the nx.graph function to create a graph with size as the x-axis and brightness as the y-axis. This will give us a visual representation of our data.

For our second research question, we will be analyzing the redshift values provided to us in the dataset. We will use a similar algorithm to order galaxies from smallest to greatest red shift values. Since redshift values are associated with the speed of growth of a certain galaxy, the algorithm will give us a better idea of which galaxies are expanding and how fast they are expanding. This part will take us an estimated five to six hours to brainstorm and implement the structure of our program. The greatest redshift value will be determined by our indexed list which would be the last value. We will use a series of if/elif statements to calculate the redshift ranges of various groups of galaxies. By doing so, we can better understand how many galaxies are moving at what rate. Through calculating the galaxy that is moving the fastest, we can compare it to a reference point which will be the Earth and find how fast the universe is expanding.

For our third research question, we also utilized a similar strategy to identify a possible correlation between the size and the redshift values of galaxies. We used the linear correlation functions to implement our graph. We also added a line to indicate the largest and smallest redshift values. For our fourth question, we created a graph showing the relationship between the redshift value and the brightness of spectral bands. We utilized the graphing function to observe this relationship.

Part Three:

We will develop and test our code by doing pair programming and dividing the role between the two team members. One member will be working on typing code and translating what is decided as a final result into syntax. The resources we will be using includes the Canopy interpreter, our dataset, and further analysis and research into galactical components if necessary to craft a better understanding of the context of our project. We will rotate roles so that both partner contribute equally to the written code and have time to review and improve the current drafts of our program.

We will develop our code by using a shared Google document and inputting ideas that we have to improve our "pseudo" plan. After crafting a solid idea, we will meet up and identify our program structure and perform pair programming to make sure that both members are equally participating in the development of the program. We will switch spots in the pair programming process so we both have experience as the coder or ideator. Most of our work will be meeting up multiple times a week to share ideas about how to best present our work. We also be visiting office hours to clarify questions we may have on the appropriate data types used and the best representation of our graph for the first research questions regarding the relationship between the brightness level and size of a galaxy. This part will take us around two to three hours, and probably some time in TA's office hours.

Part Four:

We will analyze the data that was computed and determine the likelihood of life found on other galaxies and using the speed of the redshifted light, we will determine the speed of our expanding universe. Part Four will be a detailed analysis of our results and we will write up our final findings. In addition, we will be testing out bugs by adding in assert statements and print statements to ensure that the algorithm will work for any type of value. We will also set the range of the parameters of our function. For example, we will not allow the program to take in a parameter that is a boolean or float because the galaxies should have a certain integer name associated with it already. Our test statement will test for potential inputs and outputs that may not be what we expect and edit our algorithm based off these errors. We will also reach out to classmates and TAs to gather ideas of potential bugs or improvement in the final stage of the program development. This part will take us around a day or two to receive helpful feedback and implement improvements needed as suggested by the Professor and TAs.

Evaluated Work Plan:

I think most of our work plan went pretty smoothly. However, our goals were much more ambitious than we thought. We initially wanted to craft graphs that would show the relationship between various data sets. However, we were very active on using dataset such as lists and dictionaries. Most of our data was sorted for us, because when we added the values to a list, they came in order of the rows in our csv file. However, the difficult part of our work process was to identify what kind of graph would best represent our data in the most clear and coherent way. We were pretty accurate in our work plan estimates. It took us around 12 hours at office hours to complete our rough structure of our python file. Surprisingly, our csv file had many problems in the beginning and we couldn't get the right version of the csv file to open in our extract data function. However, after converting the csv file to the right format, we were able to use our data to test our function. Our estimates were pretty good because in terms of time, we had sufficient time to complete our graph and obtain our data to work with.

8) Testing.

In our main.py file, we tested our code using assert statements. Since we could not really test the plots itself, we checked if our program was running efficiently and taking in correct data by comparing the length of each column/list to make sure they are equivalent. we are comparing each component (brightness, size, etc.) to make sure we grabbed the correct information from our dataset. We added assert statements to get_min_max_redshift to check the minimum and maximum value. We also added assert statements to galaxy_speed to make sure that all of the categories added together equal the total and that all of the categories added up to 100%. Other possible testing strategies that could have been implemented would be to use smaller test files or just one column of the data to check.

Live Presentation or Video?

B) We will be doing a 2 minute presentation about your project on Tuesday June 5, 2018 at 2:30-4:20pm.

Collaboration.

We collaborated with the TAs in office hours.