ML & Neural Networks

Maybelline Zuniga

Analyzing Galaxy Data using ML models

Abstract

This analysis aimed to explore the expanding universe by examining current galaxy data. Through this investigation, I provided evidence supporting Hubble's Law and the theory of cosmic expansion. The results revealed a strong linear relationship between a galaxy's radial velocity and redshift. Additionally, the analysis highlighted biases in the data, stemming from the limitations of current technology and the challenges in detecting light from distant galaxies. These factors may affect the accuracy of our models, but they also offer valuable insights into the universe.

Data from nearby galaxies sourced from the NASA/IPAC Extragalactic Database were preprocessed and analyzed using Python and machine learning models. The full data analysis is in a separate document and should complement this report.

Introduction

Astronomy data is vast and organizing it, cleaning it, and analyzing it can be taxing. That’s why machine learning models are convenient for handling astronomical data. Machine learning models take care of the tedious work so that people can focus on what’s important.

K-means clustering and linear regression using Python are used to analyze the data downloaded from the NASA/IPAC Extragalactic Database website. The original data file that was downloaded was in text format. It included the name of each galaxy object and its properties like RA, DEC, Type, Velocity, Redshift, etc. The file was converted to a CSV file before preprocessing and cleaning in Jupyter notebooks.

During the data analysis and results evaluation, certain properties of galaxies and light were validated. For example, the data showed the direct relationship between a galaxy's radial velocity and redshift. The data also showed the limitations and biases in the data that could lead to wrong conclusions when comparing magnitude with redshift/velocity.

Machine learning models can help mitigate these blind spots or biases in the data. Models like linear regression can help predict distances and velocities for galactic objects that are too dim to be detected by the current technology.

Background

Redshift:

Redshift can be explained using the Doppler effect. In the Doppler effect, sound waves are pushed together when a siren is moving towards the observer causing a higher frequency sound, and stretched out when moving away from the observer causing a lower frequency sound. Similarly, when a galaxy is moving toward the Earth, the light waves are compressed causing them to be blueshifted. When a galaxy is moving away from Earth, the light waves are stretched causing them to be redshifted.

K-Means Clustering (ML Libraries) :

K-means clustering organizes data into clusters without the data being labeled beforehand. This means the algorithm is unsupervised. K-means clustering can be used to bypass the tedious labeling of each value. Each cluster has a centroid or a mean and has similarities or ties to its centroid.

The data is read using pandas and then put into a data frame using numpy. The correct columns are extracted and then the data is scaled to have a mean of zero and a standard deviation of one so the large range differences don’t affect the clustering results. Then, KMeans is used to set parameters like the number of clusters and random states. The labels and centroids are defined by k means. When the data is visualized, the results show the final centroids and clusters found. The different clusters are shown in different colors using plt. scatter().

When the number of clusters is defined in KMeans(), the K-means algorithm will choose a random center or centroid and try to find similar values or values that are close to the centroid. Then the center is re-calculated or the cluster mean is re-calculated, and that becomes the new centroid. Then, the values are relabeled into new clusters close to the new centers. This is reiterated until the means of the cluster get close enough to the current centroids that the clusters stabilize or converge. In other words, the centroids don't change much between iterations.

Linear Regression (from scratch) :

Linear regression is more straightforward because it deals with linear relationships. The domain and the range are defined with x being redshifts and y being velocities. The mean of x and the mean of y are found. Then, the slope is calculated using calculus.

The numerator of the slope is the sum of all the products. Each product is the change of the redshift from the redshift mean times the change of the velocity from the velocity mean. The products for each row are added together to create the numerator of the slope.

Numerator = sum [(x - x_ mean) * (y - y_mean)]

The denominator is the sum of all the squares. Each square is the change of the redshift from the redshift mean.

Denominator = sum [(x - mean_x) ** 2]

The slope is then calculated as the numerator divided by the denominator.

The slope-intercept formula is then used to define the y-intercept.

y = mx + b

y_mean = slope * x_mean + y_intercept

b is the intercept on the y-axis

m is the slope

x is the mean of redshifts

y is the mean of the velocities

-> y_intercept = y_mean - (slope * x_mean)

The y-intercept is used to predict more velocities by inputting redshift values, x.

predicted_y = slope * x + y_intercept

The data is visualized in a scatterplot using plt. scatter() with x and y which are the redshifts and velocities. The regression line is plotted with a different color to show the predicted velocities from the redshifts in the data.

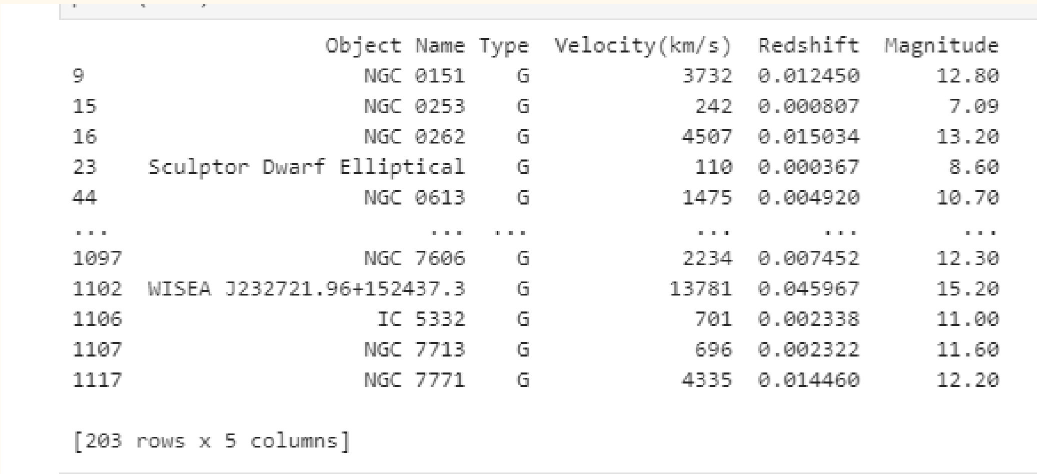
Methodology

1. Research Public Data
   1. Data was downloaded from the NASA/IPAC Extragalactic Database in text format
   2. The text file was converted to a CSV file
2. Prepare Data
   1. Cleaned Data using python in jupyter notebook
   2. Normalized data in the notebook
3. Analyze Data
   1. Created histograms and scatterplots using matplotlib
4. Implement ML Algorithms
   1. K-means using ml libraries
   2. Linear regression from scratch
5. Models
   1. Show results or visuals
6. Analyze results
   1. Find patterns
   2. Interpret findings
   3. Validate Hubble's Law
7. Limitations and Bias
   1. Compare both algorithms
   2. Discuss performance

Results

1. Prepare Data

After the data was cleaned this was the result

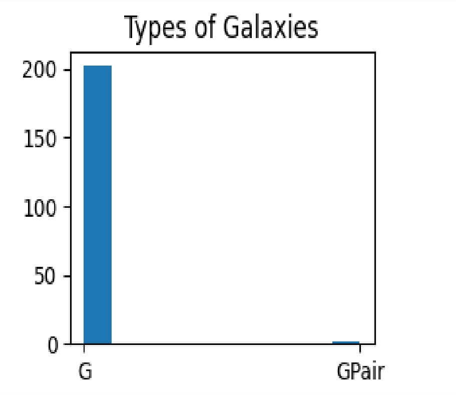


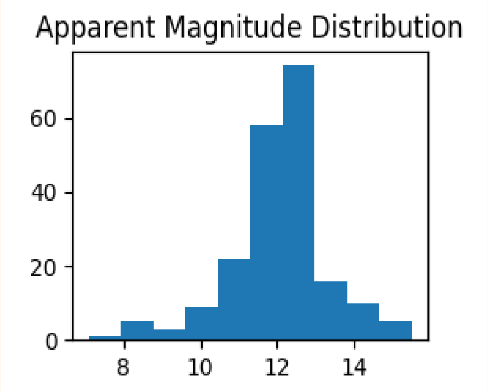
```
                   Object Name Type  Velocity(km/s)  Redshift  Magnitude
9                     NGC 0151    G            3732  0.012450      12.80
15                    NGC 0253    G             242  0.000807       7.09
16                    NGC 0262    G            4507  0.015034      13.20
23    Sculptor Dwarf Elliptical    G             110  0.000367       8.60
44                    NGC 0613    G            1475  0.004920      10.70
...                        ...  ...             ...       ...        ...
1097                  NGC 7606    G            2234  0.007452      12.30
1102  WISEA J232721.96+152437.3    G           13781  0.045967      15.20
1106                   IC 5332    G             701  0.002338      11.00
1107                  NGC 7713    G             696  0.002322      11.60
1117                  NGC 7771    G            4335  0.014460      12.20

[203 rows x 5 columns]
```

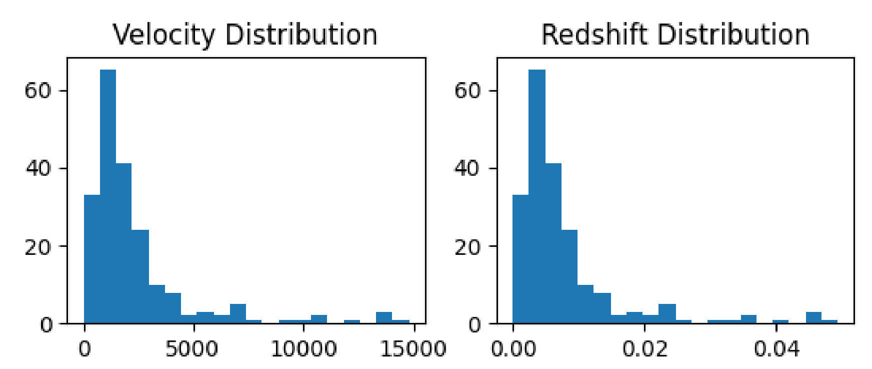
8. Analyze Data

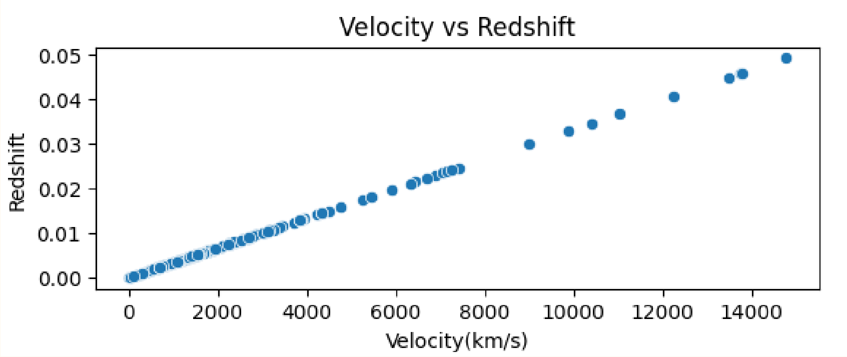
Visuals from analyzing data

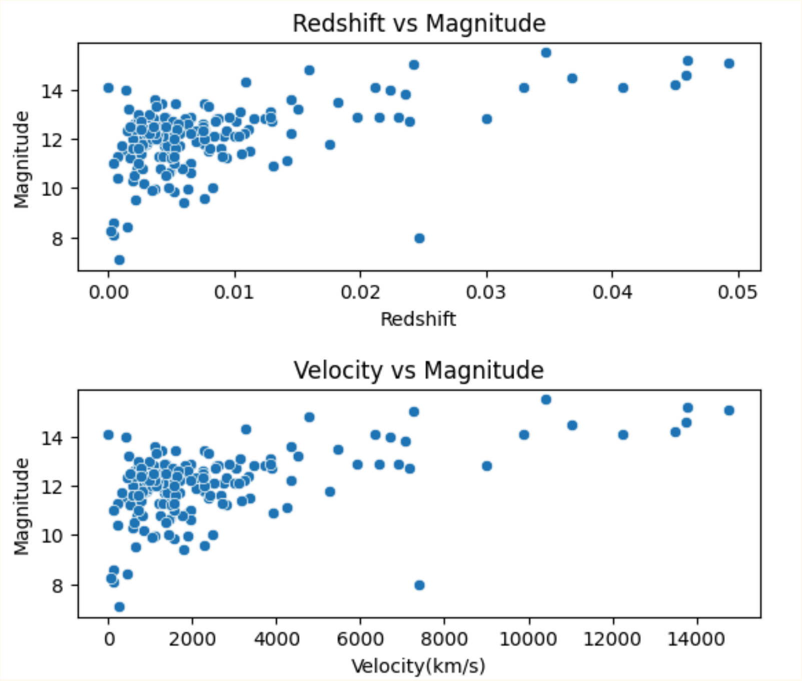




Description: Most galaxies in the data set are spiral galaxies while only very few are galaxy pairs. An apparent magnitude of 1 is way brighter than 14, very dim light

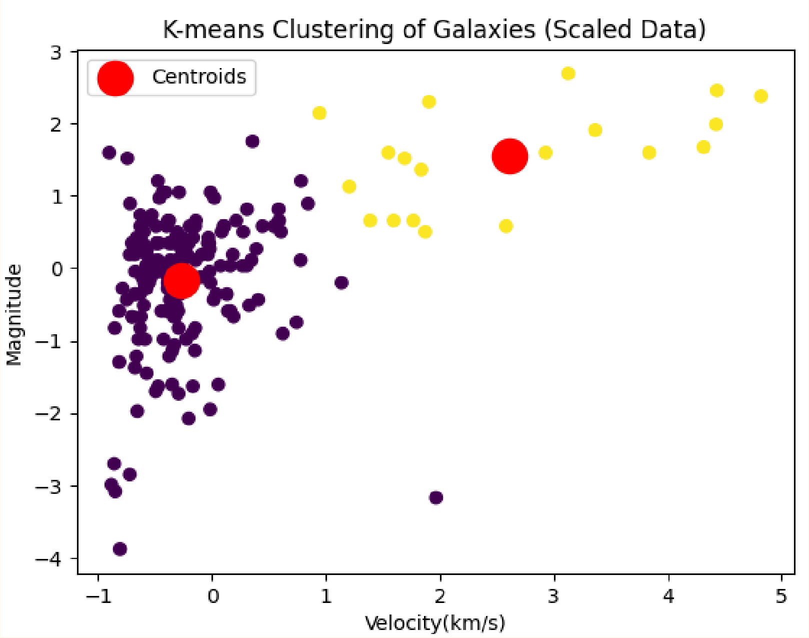


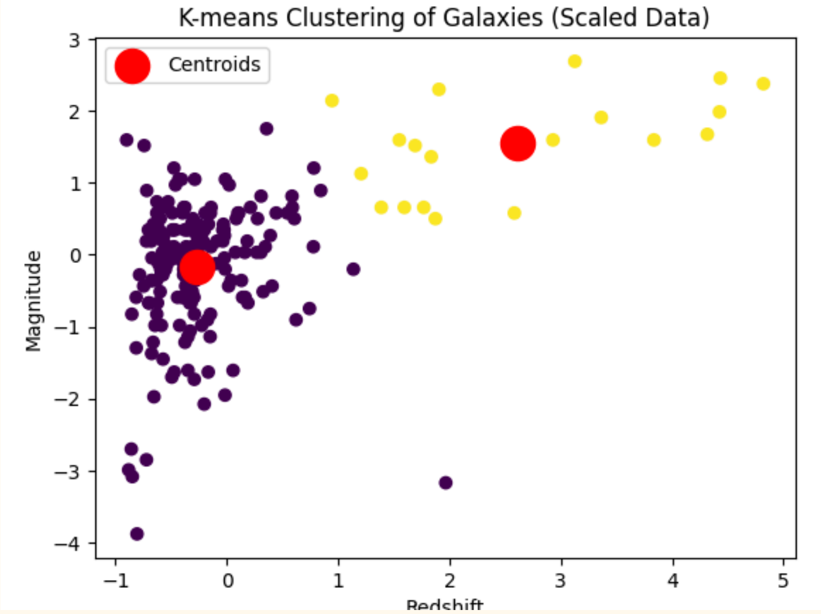




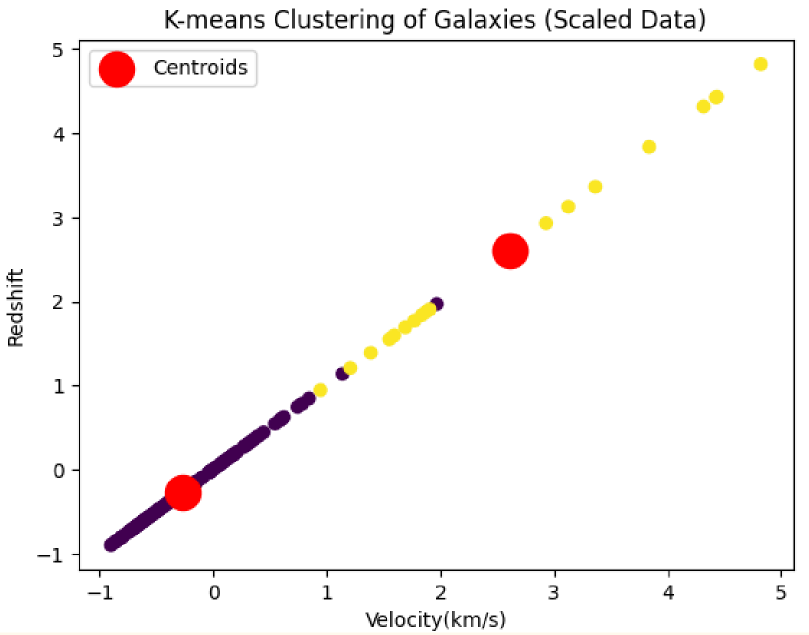
Description: A directly proportional relationship was found between heliocentric radial velocity and redshift. Clusters can be seen to gather around certain redshift or velocity.

9. Implement ML Algorithms

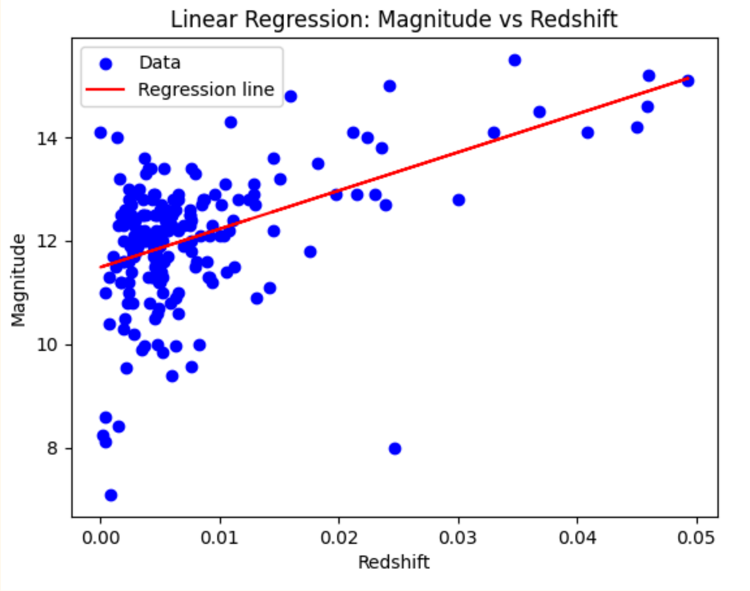




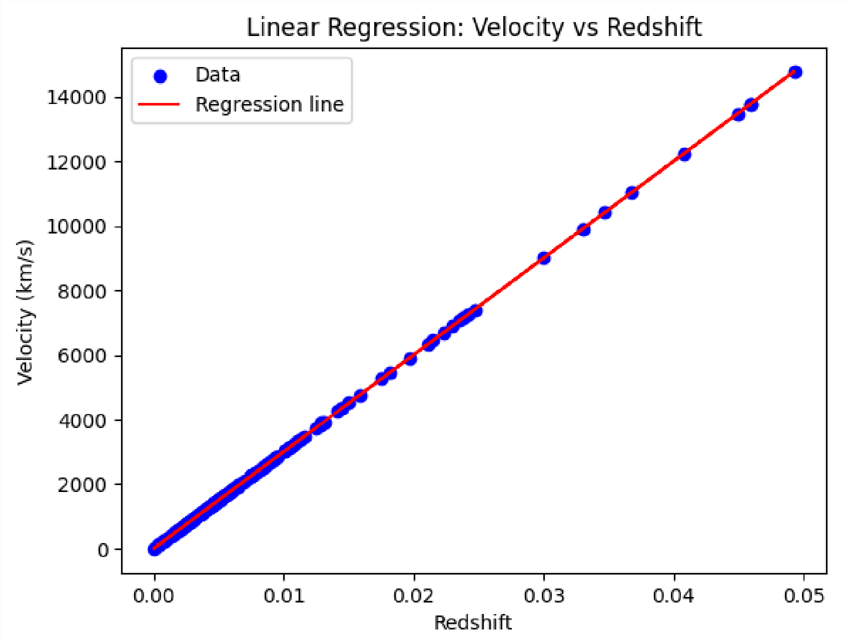
Description: 2 clusters shown. Galaxies that are dimmer tend to have a higher redshift. This means they are moving away from Earth at a higher velocity. Most galaxies found are in lower redshifts because they are closer to Earth.



Description: showing a linear relationship between redshift and radial velocity. The scaled data shows that the average is set to 0. The higher the redshift, the higher the radial velocity



Description: The higher the redshift, the dimmer the galaxy. This means they are farther away. Closer galaxies will have lower redshifts and a greater diversity of apparent magnitudes or brightness.



Description: A linear regression line shows predicted velocities. The slope is the constant or coefficient that can be used to approximate velocity. The slope found was 299795.4596774723 km/s which is approximately the speed of light, 299792.458 km / s.

Validation

The slope found during linear regression shows the **Doppler effect** where

v = c x z

- **v = radial velocity km/s**
- **c = speed of light 299792.458 km / s**
- **z =∆ λ/λ , redshift**

Linear regression can be used to predict Hubble’s constant just like we showed the Doppler effect above. The distance of the galaxies from Earth can be calculated using several methods. Hubble used the distances and the radial velocities to find the regression line or the slope connecting both variables. Now, a galaxy’s radial velocity or distance can be calculated using just Hubble’s constant.

Hubble's Law

v = H x d or c x z = H x d

- **v = radial velocity km/s**
- **H =Hubble constant, approximately 70 km/s/Mpc**
- **d = distance to the object, megaparsecs (Mpc).**

Ex) Galaxy : NGC 0151

v = 3732 km/s

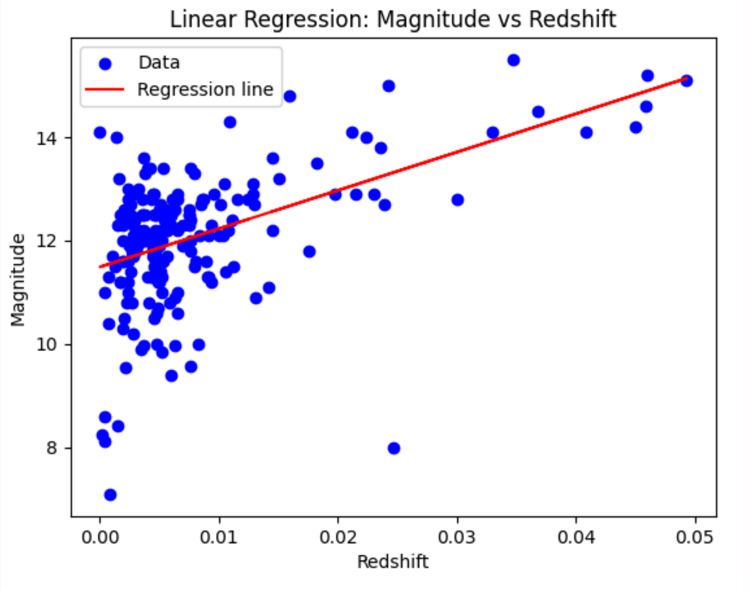
H = 70 km/s/Mpc

d = ?

d = v / H = (3732 km/s) / (70 km/s/Mpc) = 52 Mpc away from Earth

Limitations

The data gathered by the current technology is limited due to sensitivity to light and the effects of redshift and this limitation gives us incomplete data.



There are some important truths about the data that cannot be dismissed.

1. There is more diversity in apparent magnitude at lower redshifts(closer galaxies).
2. There is less diversity in apparent magnitude at higher redshifts (farther galaxies).
3. Due to the inverse square law, galaxies with higher luminosity or intrinsic brightness are easier to detect while galaxies with a lower luminosity are harder to detect.
4. The more redshifted or stretched the light gets, the harder it is to detect, and even harder if the galaxy has low luminosity.
5. However, galaxies with higher redshift AND higher luminosity(intrinsic brightness) are more likely to be detected.
6. Galaxies with lower luminosity AND higher redshifts are likely to be undetectable.

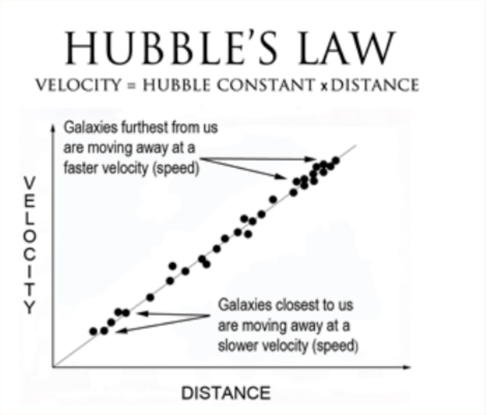
Considering this, it makes sense that the galaxies detected with the highest redshifts are less numerous in the graph. Diversity decreases significantly due to these constraints from our current technology, the inverse square law, and the effects of redshift on light. The galaxies that make it past the threshold of detectability from Earth are galaxies with the highest luminosity which shows a **bias** in the data. This could lead to discrepancies in Hubble's constant and the current model of the universe.

Evaluation

The data around redshifts supports the theory that the universe is expanding. As the galaxies get farther and farther away, the more redshifted (z) or the faster their velocity since

v = z * c.

Hubble's Law: v = H x d



Machine learning models like linear regression are useful in astronomy. The regression line between distance and redshift (distance versus z*c) shows the universe expanding.

To go further with the data, machine learning models can be used to predict the existence of undetected galaxies yet to be discovered. It could be used to test Hubble's Law at the far edges of the universe.

Conclusions

The aim was to gain insight into the universe by analyzing current data on galaxies. I showed supporting evidence for Hubble's Law and the expanding universe. The results show the linear relationship between radial velocity and redshift, the bias in the data due to limitations in our technology, and the challenges of detecting light at great distances.

References

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#examples-using-sklearn-cluster-kmeans

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

https://www.coursera.org/learn/uol-cm3015-machine-learning-and-neural-networks/lecture/THatt/3-102-linear-regression

https://www.coursera.org/learn/uol-cm3015-machine-learning-and-neural-networks/lecture/uSQf3/6-102-clustering

https://www.coursera.org/learn/uol-cm3005-data-science/lecture/WekhK/2-201-series-and-data-frames-in-pandas

https://www.coursera.org/learn/uol-cm3005-data-science/lecture/bSYqs/2-001-introduction-to-numpy

https://ned.ipac.caltech.edu/byparams

https://en.wikipedia.org/wiki/NGC_151

https://astrobites.org/2016/04/20/conflicts-between-expansion-history-of-the-local-and-distant-universe/

https://docs.python.org/3/library/index.html