Final Report : Capstone Project

Aarya Bhardawaj, Rafaye Khan, Zaid Tayeh

Abstract

Galaxies are complex systems in space which us humans have just begun exploring. From black holes, transients, and even properties of galaxies such as shape and color, we are continuously discovering new information for the countless galaxies in space. We will be analyzing multiple datasets to answer 3 interesting research questions. We will use t test, hypothesis test, bootstrapping, and an ANOVA test to analyze the datasets and come to a conclusion based on the results. We will be determining how the mean spectra for each galaxy differ, whether there are equal fractions of mwebv, the foreground extinction due to intervening cosmic dust, across all supernova classes in the training and testing data, and if the shape of the galaxy (elliptical or spiral) correlates with its spectra (color). For the first research question, we were able to conclude that the average spectra for each galaxy was different to one another with some galaxies having a much higher spectra compared to the others. For the second research question we were able to conclude that there are not equal fractions of mwebv scroll all supernova classes. For the last question we were able to conclude that
With these given conclusions, in the future we can determine as to why our conclusions resulted the way they did and determine the reason why we could not find correlation between our research parameters.

First Research Question

Introduction

Understanding galaxies' spectra is essential for modern astronomy research into the composition, age, and development of galaxies. Spectra are the patterns of light produced by astronomical objects, and they provide important information about the physical properties of these objects (National Aeronautics and Space Administration [NASA]). How do the mean spectra of each galaxy differ is a key subject in this report. is researched in R and through data visualization. In this report the focus is on understanding how the mean spectra of each galaxy differ. The data used for this analysis was obtained from the Galaxy Zoo dataset (Lintott el al.2008). In order to visually represent the differences between each galaxy, plots were made using the mean spectra of six different galaxies, which were then extracted and analyzed. Insights concerning the makeup of galaxies, their evolutionary history, and how their spectra can be used to advance our understanding of the universe are sought after by the report. To give a thorough and aesthetically pleasing analysis of the data was the goal of this study question. The goal of this study question was to provide a thorough and aesthetically pleasing analysis of the data, the insights gained from this analysis can help further our understanding of the makeup of galaxies.

Data

In order to answer this question. Data from the Galaxy Zoo dataset was needed in order to complete the visualizations. The Galaxy Zoo dataset contains an extensive collection of images and classifications for different galaxies and is a valuable resource used in order to gain a better understanding of the universe. The dataset is useful as it has the inclusion of the mean spectra for each galaxy which provides valuable insight. By extracting the mean spectra values from each column of the table, I was able to create visualizations that clearly demonstrated the differences between the mean spectra for each galaxy. The Galaxy Zoo dataset and the analysis overall is important for understanding the properties of galaxies.

Methods

The dataset used in order to answer the question of "How does the Mean Spectra for each Galaxy differ? More specifically on the GalaxyZoo site, the data which contained the classifications for galaxies as well as their spectra which was table 2. By using R the mean value of the spectra for each galaxy was extracted and then could be used for the data visualization. Each galaxy was listed in the excel table 2 file and in each column the mean spectra for each galaxy was listed. Aim of the hypothesis test (figure 7) is to perform a one sample T-Test is performed on Data in order to determine whether the mean of the population (Average spectra/brightness for each galaxy) from which the sample was drawn is significantly different from the value specified by 'mu'. The output of the test shows that the calculated t value is -4.86 with a p value of 0.000462. The p value indicates the probability of obtaining a sample mean as extreme as the observed one, assuming the null hypothesis (that the true mean is equal to 0.6) is true. A small p value suggests that the null hypothesis can be rejected in favor of the alternative hypothesis. In this case the p-value is less than 0.05 which suggests that the null hypothesis should be rejected and that the true mean of the average spectra of all the galaxies is not likely equal to 0.6. Furthermore, the sample mean of '0.0221' is given, along with a 95% confidence interval of '(0.021,0.4216'). This interval suggests that we can be 95% confident that the average spectra for each galaxy falls within this range. Due to the fact we wanted to illustrate how the average spectra for each galaxy differed, the first data visualization (figure 1) used was a bar graph as it allowed us to easily compare the average spectra for multiple galaxies at once. Bar graphs are ideal for displaying quantitative data for the purpose of showing comparisons between different groups and categories. In this case, along the x-axis each galaxy was listed clearly, while on the y-axis the quantitative variables of the average/mean spectra were listed. We decided to use a graphical data visualization with a bar chart as we felt this was the most effective way to communicate our results to an audience as our results contained both quantitative and qualitative data. For our **Second Visualization (figure 2)**, in order to emphasize the distribution aspect of our question to our viewers we decided to utilize a density plot which contained the same data in our bar graph. In order to properly illustrate the distribution of the average spectra across the different galaxies, we felt a density plot would effectively present the

distribution of the spectral characteristics across multiple galaxies in the Galaxy Zoo Dataset. Although the first visualization shows the mean spectra for each galaxy this visualization focuses more heavily on the actual distribution between galaxies.

Results

Modern astronomy study places a strong emphasis on understanding galaxies' spectra because they include vital details on the chemical make-up, age, and development of galaxies. This field places a lot of emphasis on the investigation of mean spectra for various galaxies since it gives astronomers insight into the distinctive characteristics of each galaxy. This paper looked into how the average spectra of six different galaxies differ from one another using R and data visualization tools. The findings of this investigation can throw light on the underlying physical mechanisms causing these variations, which in turn can reveal important details about the creation and development of galaxies. This paper also aimed to give a thorough and aesthetically pleasing analysis of the data, emphasizing the importance of spectroscopic data in increasing our comprehension of the cosmos. This report makes a contribution to the continuous quest to unravel the cosmos' mysteries and increase our understanding of the universe by diving into the subtleties of galaxy spectra.

Second Research Question

Introduction

Transients refer to astronomical phenomena with durations of fractions of a second to weeks or years. Typically they are extreme, short-lived events associated with the total or partial destruction of an astrophysical object. Supernovae are types of many transients (*Nature news*). Supernovae are powerful and luminous explosions of stars. These fascinating explosions occur during the last evolutionary stages of a massive star or when a white dwarf is triggered into runway nuclear fusion. These violent explosions of stars can be used to study the beautiful cosmos. In order to effectively study supernovae, it is important to accurately classify them into different types based on their observed characteristics. One of the key factors that affects the observed properties of a supernova is the amount of dust along the line of sight. This parameter is measured as a “mwebv” value which is the foreground “extinction” due to intervening cosmic dust that is defined in units of magnitude (*What is a supernova?* 2021). For this question, we will be determining whether there are equal fractions of mwebv across all supernova classes in the training and testing data.

Data

To analyze this question, we will observe the PLAsTiCC dataset which is available on Kaggle(*Kaggle* 2018). This dataset contains the relevant information which needs to be observed

in order to answer this question. In particular, there are a total 14 transient classes which can be accessed under the "true_target" parameter, as seen in figure 3. However for this question, only the transient classes which are of importance are the ones which are caused by supernovae. Therefore, only the transient classes of 90, 67, 52, 42, 62, and 95 are the ones which need to be analyzed. Before any statistical methods are performed on the dataset, it is important to first filter out the data appropriately so that it only includes these 6 transient classes. To perform this data wrangling, we first create a new variable to save the filtered data under and then use the filter function on the raw training dataset downloaded from Kaggle to only include information on transients which are of the 6 supernovae classes. After filtering the dataset so it only includes the 6 transient classes of interest, figure 4 shows the new mutated data in the form of a barplot. Each bar represents the respective supernova class shown by the x axis. The height of each bar represents the number of instances in the dataset that belong to that class. Figure 5 shows the mwebv values for each respective supernovae class in the form of a boxplot. This visualization shows how the mwebv values are spread out in the 6 supernovae classes before any statistical methods are applied. This is the final dataset that will be used for the rest of the analysis.

Methods

The methods used to solve this question are bootstrapping and an ANOVA, analysis of variance. These methods were all implemented in R. First, we took the filtered dataset and created 1000 resampled datasets using the "sample_n()" function in R within a for loop. This is random sampling with replacement on the filtered dataset. These resampled datasets were all stored in a new variable called "resampled_datasets". Since we are trying to determine whether each supernova class has equal amounts of mwebv, we need to consider the mean mwebv values for each class. Figure 6 shows the mean mwebv values for each respective supernovae class in the form of a boxplot. Note that this plot is from the resampled data. Next before performing an ANOVA test, we combined all resampled datasets and grouped them by their class using the "group_by" function in R. We calculate the mean for all of these classes and store them into a new variable. Before we can perform an ANOVA test on this data we must check if the residuals are evenly distributed first. This is to make sure the assumptions are satisfied before performing an ANOVA test. As seen in figure 7, the histogram shows that the residuals are indeed evenly distributed thus it is possible to carry out the ANOVA test. Now we have the F - statistics on the mean mwebv values for each class. After the ANOVA test we simply extract all the bootstrap distributions from all the classes and compute the confidence intervals.

Results

Based on the analysis, we can conclude that there is a significant difference in the mean mwebv values across the different supernova classes in the PLAsTiCC dataset. The ANOVA test indicates that the observed differences in mean mwebv values between the supernova classes are statistically significant ($p < 0.001$). Additionally, the confidence intervals computed from the

bootstrap distribution suggest that the difference in mean mwebv values between at least two of the supernova classes is likely to be meaningful and not due to chance.

The ANOVA table shows the degrees of freedom (Df), the sum of squares (Sum Sq), the mean square (Mean Sq), the F statistic (F value), and the corresponding p-value (Pr(>F)).
The null hypothesis for ANOVA is that there is no significant difference between the mean mwebv values of the different supernova classes, and the alternative hypothesis is that there is at least one group with a significantly different mean mwebv value. The tabular summary shows that the p-value is less than 0.001 (Pr(>F) = 7.891e-05), indicating strong evidence against the null hypothesis. Therefore, we can conclude that there is a significant difference between the mean mwebv values of the different supernova classes.

The confidence interval obtained ranges from 0.0003043015 to 2.246597. Note that this is a 95% confidence interval so if we were to repeat this resampling procedure many times, approximately 95% of the intervals generated would contain the true population parameter. This confidence interval suggests that the true difference in mean mwebv values between the different supernova classes is likely to be between 0.0003 and 2.2466. With this we can safely reject the null as we know that the difference of mean within the mwebv values are evident.

Data Visualizations

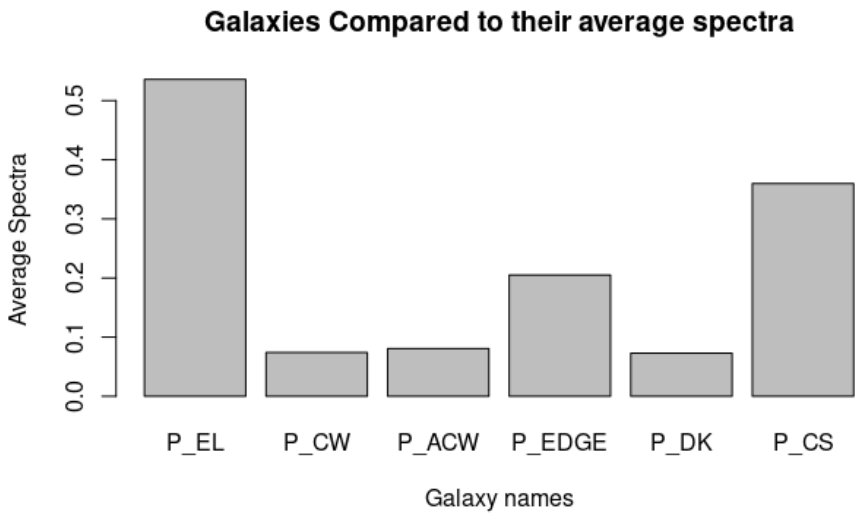


Figure 1

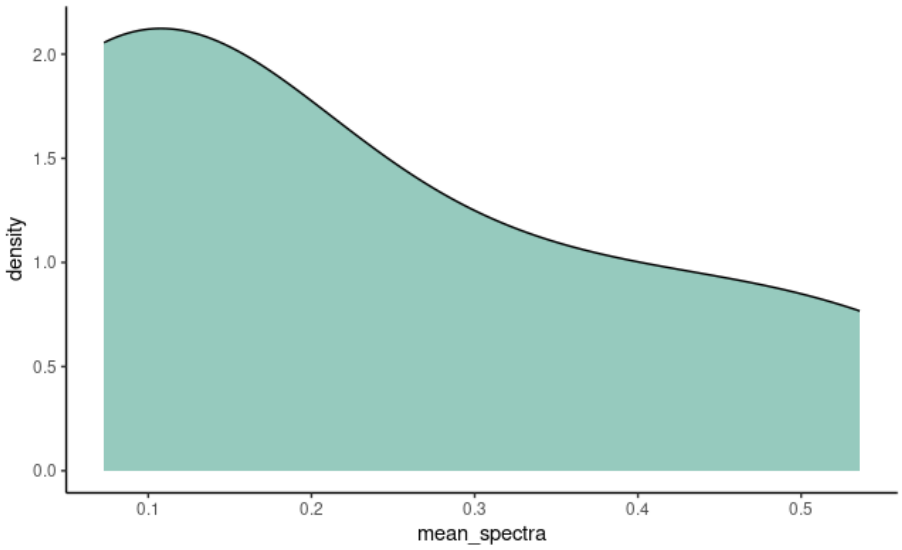


Figure 2

model class num[a]: name	model description	contributor(s)[b]
90: SNIa	WD detonation, Type Ia SN	RK
67: SNIa-91bg	Peculiar type Ia: 91bg	SG,LG
52: SNIax	Peculiar SNIax	SJ,MD
42: SNII	Core Collapse, Type II SN	SG,LG:RK,JRP:VAV
62: SNIbc	Core Collapse, Type Ibc SN	VAV:RK,JRP
95: SLSN-I	Super-Lum. SN (magnetar)	VAV
15: TDE	Tidal Disruption Event	VAV
64: KN	Kilonova (NS-NS merger)	DK,GN
88: AGN	Active Galactic Nuclei	SD
92: RRL	RR lyrae	SD
65: M-dwarf	M-dwarf stellar flare	SD
16: EB	Eclipsing Binary stars	AP
53: Mira	Pulsating variable stars	RH
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN

Figure 3

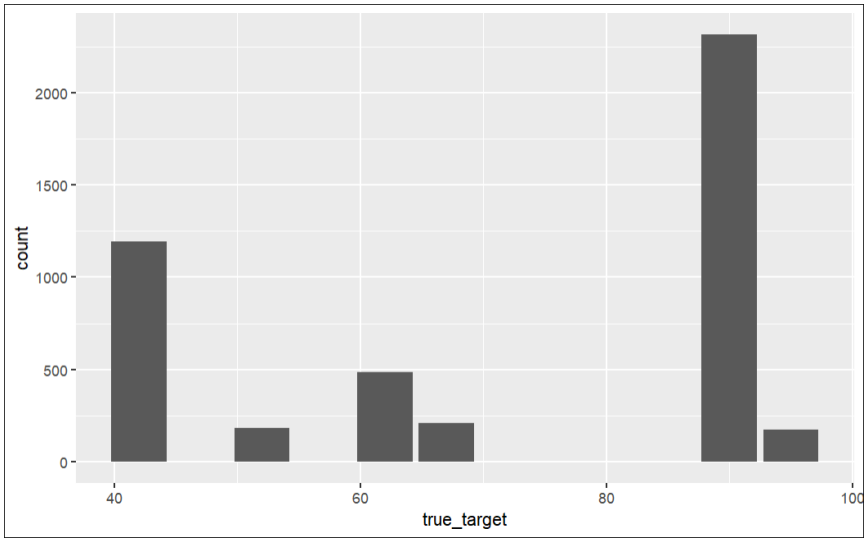


Figure 4

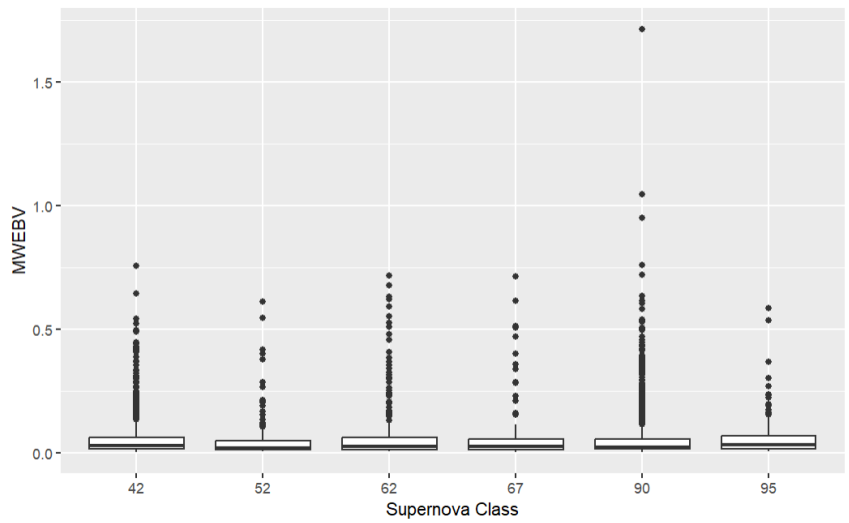


Figure 5

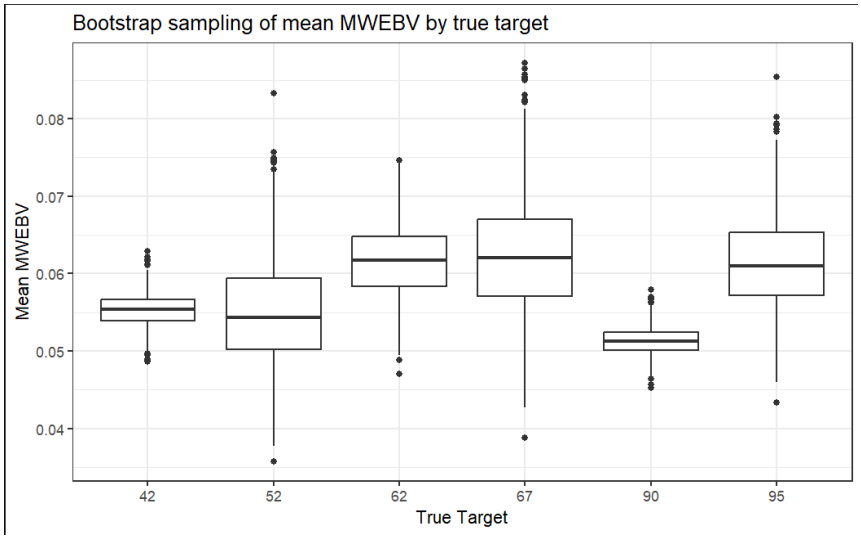


Figure 6

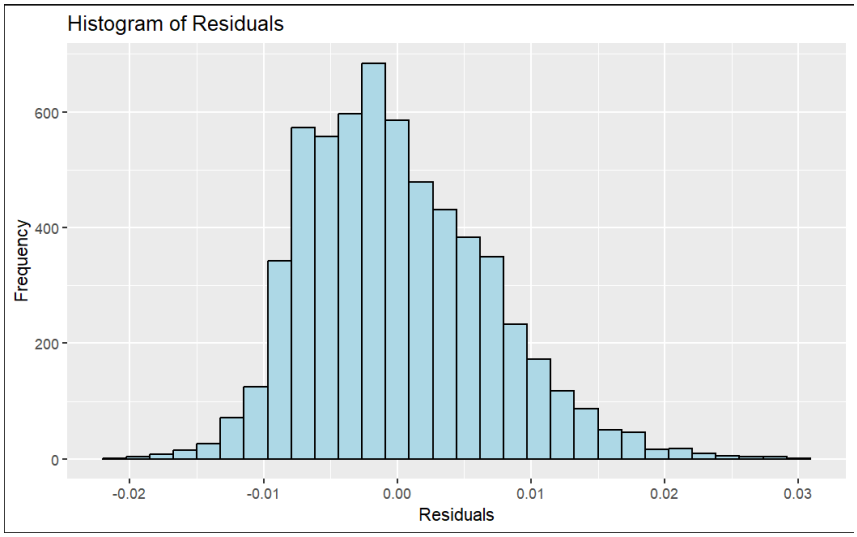


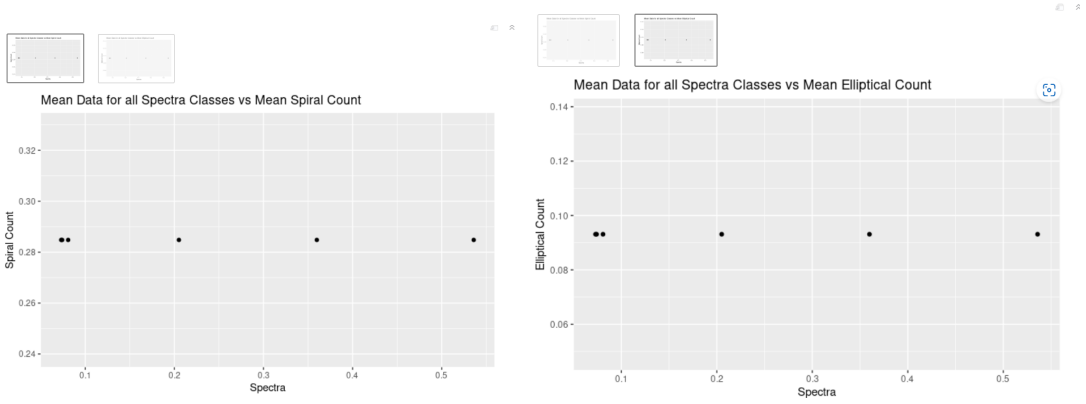
Figure 7

```
> t.test(dayta, mu=0.6)

        One Sample t-test

data:  dayta
t = -4.8614, df = 5, p-value = 0.004627
alternative hypothesis: true mean is not equal to 0.6
95 percent confidence interval:
 0.02124017 0.42160356
sample estimates:
mean of x
0.2214219
```

Figure 8



Third Research Question

Introduction

For this research question, we analyzed whether the shape of a galaxy correlates with its spectra. We analyzed two shapes (Elliptical and Spiral) and used all the data from table 2 of the Galaxy Zoo dataset. We set out to find whether galaxies of different shapes emit different spectra (color). If they do, that could have told astronomers a lot about the inner working relationships between two seemingly unrelated galactic phenomena. We analyzed 6 different classes of galaxies and analyzed all the spectra data points and created two visualizations and a hypothesis test for the correlation to the data we had regarding the shape. This information will graph the exact data points for mean spectra versus the binary data for whether it is spiral or elliptical. It will allow us to create a hypothesis test (null hypothesis and alternate hypothesis) for our research question and then simulate it which will give us information on whether to reject or fail to reject the null hypothesis.

Data

We obtained all the data we needed from the Galaxy Zoo dataset, specifically using table 2 for this research question. We used the information listed about spectra (listed underneath each class of galaxy) and correlated it with the binary information for the shape of the galaxy (listed underneath the Spiral or Elliptical columns). We arranged the data in the spiral and elliptical class in accordance with each value for spectra as it would make the data look more presentable and we would be able to create an in depth analysis for each following visualization.

Methods

The question of whether the shape of a galaxy correlates with its spectra can be answered with a manipulation of data and a scatter point visualization. Hence through R, we graphed the mean spectra for each class; we took the mean of all the spectra data for each class measured. We then took the mean count for the shape data on elliptical and spiral galaxies and graphed two scatterplots, one for the Spiral class and one for the elliptical. We also arranged the y variable (shape data) so that each data point aligned with the correct class. Hence, we obtained two scatter plots that show this correlation. For the Hypothesis test, we set the null hypothesis to be that there is no correlation between spectra and the shape of a galaxy. Hence the obtained p-values would be lower than 0.05 (our significance level). The alternate hypothesis was that there is a correlation. We set up the test with the two vectors containing the mean spectra value (x) for each class of galaxy and then the related variable (y), the likelihood of a galaxy being both spiral and elliptical.

Results

Simply by observing the observation, we found that there was no correlation between the shape of a galaxy and its spectra. However, the hypothesis test confirmed that. We found that for all mean values of spectra for each class, the mean elliptical and spiral data was the same individually. We found that spiral shaped galaxies seemed to be the most dominant shape, however, for each class, there were the same amount. The hypothesis test hence found no correlation or p-value as there was no standard deviation. This confirms our null hypothesis and we can fail to reject it. The visualization also confirmed this as for each data point in each class, it reached the same avg count value for all classes.

Discussion

For the first research question….

The discovery that the average spectra of all galaxies are distributed widely and unevenly offers crucial new information about galaxies and their makeup. The fact that some galaxies have significantly greater spectra than others may indicate that the distribution of the underlying data is not uniform and that certain galaxies have a distinct composition or evolutionary history. It's crucial to take into account the statistical significance of this finding, and the use of t tests, hypothesis testing, and ANOVA tests helped to make sure the results are accurate. Understanding how to interpret these results will help us better comprehend how galaxies evolve and change through time, as well as how their compositions vary from one another. These discoveries may also result in future

When determining whether there are equal fractions of mwebv across all supernova classes in the training and testing data, our results indicated that there was a significant difference in mwebv values between the 6 different supernovae classes. This is evident by our computed p value being less than 0.0001 in the ANOVA test and the confidence interval not including 0. However, it is possible that there may be a bias in the mwebv values across different supernova classes. Further investigation into the reasons for this difference could be useful in better understanding the underlying data and improving the accuracy of predictions. One potential avenue for future research could be to examine the effect of other variables, such as redshift or host galaxy type, on mwebv values across different supernova classes. The methods used for this question, bootstrapping and ANOVA test, are indeed useful. However, it is important to remember the assumption we used in order to carry out the analysis. We made sure that our assumptions were accurate before performing an ANOVA test, as we must verify that the groups being compared have equal variances and that the observations within each group are independent and normally distributed before we can even carry out the test. These assumptions are important to ensure the validity of the F-test used in ANOVA to determine if there are significant differences between the means of the groups being compared. In bootstrapping, we assume that the resampled

datasets are representative of the underlying population. This assumption is based on the idea that the resampled datasets are constructed by randomly sampling from the original dataset with replacement, which preserves the original data structure and distribution. However, if the original dataset has certain biases or limitations, these may be carried over into the resampled datasets and affect the results of the analysis.

For the third research question, we were limited by the uncertain data column for the shapes of the galaxies. There was quite a bit of data there that we could not make an inference about and therefore, that definitely affects the results. Also throughout the analysis, we noticed the average shape values did not differ throughout each class. However, this was not enough to draw a conclusion. We needed to conduct visualizations and a final hypothesis test to determine whether the shape of a galaxy affects its spectra. There was also no data determining longevity, hence we did not include changes in spectra as all the data was measured at a certain time stamp. This may have affected our results as the conclusion of results did not include a firm answer with such factors included.

Conclusion

In conclusion, we have improved our understanding of the complexity of galaxies by the study of many datasets using various statistical tests and visualization methods. We came to the conclusion that there are substantial disparities in the average spectra between galaxies after looking at the mean spectra for each galaxy. We also learned more about the nature of supernovae from our results about how foreground extinction from intervening cosmic dust varied across different supernova types. Also, there was no discernible connection between the form of a galaxy's spectrum and its shape according to our visualizations and research. Although these results are insightful, more study and analysis are required to properly understand the subtleties of the galaxies with their specific properties. With continued research and exploration, we can continue to expand our knowledge of the universe and the complex systems that exist within it.

Citations

Lintott, C. J., Schawinski,. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389(3), 1179–1189. https://doi.org/10.1111/j.1365-2966.2008.13689.x

National Aeronautics and Space Administration (NASA). (n.d.). Spectra of Science: Visible and Beyond. Imagine the Universe! Science Center. https://imagine.gsfc.nasa.gov/science/toolbox/spectra1.html

PLAsTiCC. (n.d.). *PLASTICC astronomical classification*. Kaggle. Retrieved April 8, 2023, from https://www.kaggle.com/competitions/PLAsTiCC-2018/data

NASA. (2021, July 23). *What is a supernova?* NASA. Retrieved April 8, 2023, from https://spaceplace.nasa.gov/supernova/en/

Nature Publishing Group. (n.d.). Nature news. Retrieved April 8, 2023, from https://www.nature.com/subjects/transient-astrophysical-phenomena#:~:text=Definition,destruction%20of%20an%20astrophysical%20object.

Contributions:

First Question and subsequent Data Visualizations, methods and data: Rafaye Khan

Second Question and subsequent Data visualizations, methods and data: Aarya Bhardawaj

Third question and subsequent Data visualizations methods and data: Zaid Tayeh

Conclusion: Rafaye and Zaid Discussion: Aarya, Rafaye and Zaid Abstract: Zaid and Aarya

One of the biggest challenges I encountered during my Capstone Project was related to group work. Coordinating schedules and ensuring that every member of the group was making progress on their assigned tasks proved to be a formidable task. However, this experience taught me valuable leadership skills, as I took the initiative to organize group meetings and progress checks. Furthermore, when one of our group members dropped the course after we had submitted the progress report, it forced us to re-evaluate the project timeline and distribution of work. Despite the challenges of group work, I am proud of the final project that we produced as a team. Collaborating with others allowed me to broaden my perspective and learn new ways of approaching complex problems. Overall, the experience was incredibly rewarding, and it has given me a newfound appreciation for the importance of teamwork and effective communication in achieving shared goals.